amse
école d'économie d'aix-marseille
aix-marseille school of economics

Working Papers / Documents de travail

Parametric models of income distributions integrating misreporting and non-response mechanisms

Mathias Silva

WP 2023 Nr 11

Parametric models of income distributions integrating misreporting and non-response mechanisms *

Mathias Silva†

Aix Marseille Univ, CNRS, AMSE, Marseille, France

This version: May 4, 2023

Abstract

Several representativeness issues affect the available data sources in studying populations' income distributions. High-income under-reporting and non-response issues have been evidenced to be particularly significant in the literature, due to their consequence in under-estimating income growth and inequality. This paper bridges several past parametric modelling attempts to account for high-income data issues in making parametric inference on income distributions at the population level. A unified parametric framework integrating parametric income distribution models and popular data replacing and reweighting corrections is developped. To exploit this framework for empirical analysis, an Approximate Bayesian Computation approach is developped. This approach updates prior beliefs on the population income distribution and the high-income data issues pressumably affecting the available data by attempting to reproduce the observed income distribution under simulations from the parametric model. Applications on simulated and EU-SILC data illustrate the performance of the approach in studying population-level mean incomes and inequality from data potentially affected by these high-income issues.

Keywords: 'Missing rich', GB2, Bayesian Inference.

JEL Code: D31, C18, C11

*I am grateful to Michel Lubrano, Stephen Bazen, Emmanuel Flachaire, Philippe Van Kerm, Markus Jäntti, Jonathan Goupille-Lebret, participants at the 8th European User Conference for EU-Microdata (GESIS, Eurostat), and seminar participants at the AMSE PhD seminar and ENS Lyon CERGIC Graduate seminar for helpful contributions and comments on earlier versions of this paper.

†Corresponding author: mathias.silva-vazquez@univ-amu.fr. Aix Marseille Univ, CNRS, AMSE, Marseille, France. AMSE - Aix-Marseille Université 5-9 Boulevard Bourdet 13001 Marseille, France. The project leading to this publication has received funding from the French government under the "France 2030" investment plan managed by the French National Research Agency (reference: ANR-17-EURE-0020) and from Excellence Initiative of Aix-Marseille University - A*MIDEX. Declaration of interest: None. All `R` replication code for the results shown in this paper are available upon request and will soon be made available on the author's Github site.

1 Introduction

The recent literature on income inequality has paid increasing attention to the dynamics and the measurement of top incomes (e.g., Atkinson and Piketty 2007, Leigh 2009, Atkinson et al. 2011, Burkhauser et al. 2017). The slowly-rising availability of tax data for research purposes along with findings concerning the recent rises in the share of incomes accumulated at the very top quantiles of the distribution (e.g., Lakner and Milanovic 2016, Alvaredo et al. 2018) have jointly brought forward the multiple deficiencies affecting the typical methods and data sources used to study income distributions.

A robustly evidenced shortcoming of these conventional approaches involves the limited quality of publicly-available household survey data, the most commonly used data source on the subject, in capturing the magnitude and trends of the income shares of the highest incomes in their population (e.g., Deaton 2005, Burdín et al. 2014, Jenkins 2017, Higgins et al. 2018, Lustig 2019).

Typically these measurement and coverage issues around the upper tail of a population's income distribution imply non-random missing information in the data (i.e., the errors are more likely or larger in magnitude for higher incomes) and therefore induce bias into any resulting distributional estimate. When ignored, this can have many clear policy implications as it leads to underestimation of income growth and inequality at the population level, along with a biased reading of their relationship and dynamics. This has motivated a vast literature on correction and estimation methods to overcome this data issue for the study of income distributions.

An important implication of this almost universal problem of missing or misreported high incomes is that any empirical strategy seeking to overcome it requires a decision on the magnitude and distribution of such errors affecting the data. As put forward by Bourguignon (2018), adjusting for measurement errors on high incomes requires a value for some or all of three key parameters: the income level beyond which measurement errors are to be corrected, the true population share of incomes above this level, and the share of undercovered population incomes.

Although external data sources can be instrumentally used to formulate informative choices for these parameters (e.g., Atkinson and Piketty 2007, Chapter 2, Bustos 2015, Blanchet et al. 2018, Jorda and Niño-Zarazúa 2019, Flachaire et al. 2022), correcting for measurement or coverage errors on high incomes is conditioned by the uncertainty around them. Broadly speaking, the precision with which inference can be made on a population's income distribution depends on the uncertainty around the form and magnitude of measurement or coverage errors affecting the available data.

This paper proposes a new empirical strategy bridging several previous results in the income inequality literature. Firstly, a parametric modelling approach is developped in the interest of integrating within a single framework all assumptions about the form of the population's income distribution and the form of the measurement or coverage issues affecting the available data. This parametric framework allows for exploiting several previously explored parametric corrections for high incomes data issues in making inference at the population level.

Secondly, a Bayesian estimation strategy allows for inference on the population's income distribution through data presumably affected by representativeness issues on the upper tail. This strategy extends the Approximate Bayesian Computation approach recently explored in Kobayashi and Kakamu (2019) and Silva (2023) in the context of income distributions. In exploiting this approach to estimate income distributions under the proposed parametric framework, the magnitudes and forms of the representativeness issues may be uncertain. Past knowledge on the possible nature of these under similar settings poses information that may be used in dealing with this uncertainty through the use of informative prior beliefs.

Finally, several applications over simulated and household survey data from the European Union's Statistics on Income and Living Conditions (EU-SILC) illustrate the performance of the proposed approach in controlled and observational settings. These applications evidence the several biases that hinder making inference on a population's income distribution if high-income representativeness issues affecting the available data are ignored. Additionally, the presented estimates suggest the presence of both high-income under-reporting and high-income non-response issues in selected EU-SILC samples. This results in population-level estimates of average incomes and inequality that are at higher levels and with higher uncertainty than their sample counterparts.

The following section presents an overview on the common causes and corrections for data errors on high incomes explored in the previous literature. Section 3 develops a parametric framework integrating popular forms of such data errors to parametric income distributions. The fourth section introduces an Approximate Bayesian Computation routine for inference on a population's income distribution through the proposed parametric framework and under magnitudes and forms for high-income data issues that might be uncertain. Section 5 presents simulated and EU-SILC data applications of the method under typical parametric forms. The sixth and final section of the paper presents concluding remarks with proposals for future work in studying high-income data issues through the proposed approach.

2 Dealing with 'missing rich' issues

In modelling income distributions, the use of parametric models is a standard. Some work has fruitfully explored the use of non- or semi-parametric methods for income distribution analysis (e.g., Jenkins 1995), yet there is vast evidence of parametric models fitting real data on incomes better than these alternatives in many different settings (e.g., Darvas 2019, Jorda et al. 2021).

The usual modelling step involves assuming that individuals' incomes y_i are distributed across its population following some three- (i.e., $\Theta \subseteq \mathbb{R}^3$) or four-parameter (i.e., $\Theta \subseteq \mathbb{R}^4$) distribution $y_i \sim f_y(.;\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Popular choices for $f_y(.;\boldsymbol{\theta})$ include the Generalized Beta family of distributions (e.g., see McDonald 1984, Jenkins 2009,Graf and Nedyalkova 2014, Chotikapanich et al. 2018, Jorda and Niño-Zarazúa 2019), in particular the four-parameter Generalized Beta distribution of the second kind (GB2, which is taken as illustratory reference in what follows) and the three-parameter Singh-Maddala (Burr XII)

distribution, and the Double Pareto Log-Normal distribution.

There are several virtues to the parametric approach aside from its generally good fit to real data on incomes. Of particular relevance is its flexibility with respect to the format of data available. Several estimators following parametric expressions for microdata or bracketed/grouped data from incomes following $y_i \sim f_y(.;\boldsymbol{\theta})$ are available for most distributions such as Generalized Method-of-Moments (GMM), Maximum Likelihood Estimation (MLE), or Bayesian inference methods.

A central consideration required in analysing data prone to high-income representativeness issues is that the distribution of observed incomes y_i^{Obs} will very unlikely follow the form of the population's income distribution $y_i \sim f_y(.;\boldsymbol{\theta})$. Within a parametric approach, however, the observed distribution can be derived under assumed parametric forms for the errors affecting the data. Jointly modelling the population income distribution component $f_y(.;\boldsymbol{\theta})$ and the high-income issues is an attempt at separating which aspects of the data reflect those of the population income distribution and which aspects are due to the high-income problems considered.

In describing the nature of the 'missing rich' (MR) problem, Lustig (2020) points at the many different issues affecting the upper tail of the observed income distribution in usual data sources. In the context of survey data, the main focus of this paper, one first source of observed MR may arise from noncoverage errors in the sampling design itself as a consequence of the sparseness and irreplaceability of high income households. High-income households are generally so few and so dissimilar between themselves that households on any part of the upper tail of the distribution may have a zero probability of inclusion in the achieved survey sample.

A second possible source for MR in survey data involves reporting issues either in the form of unit or item nonresponse (i.e., high-income households refusing to respond to the survey or particularly to the items concerning their income level, respectively) or in the form of under-reporting of income levels when responding to the survey. Even if the sampling scheme is designed to be representative of the income distribution of the entire population of interest, unit or item non-response may yield an achieved sample which is not and particularly so when this nonresponse occurs more significantly for households on the upper tail of the income distribution. In a similar way, with income under-reporting the achieved survey sample may yield an income distribution which is not representative of the population's true income distribution if under-reporting is particularly present for high-income households.

Finally, a third possible source can be found within the data provision procedures commonly used by the institutions in charge of distributing publicly-available household survey datasets. In the interest of statistical disclosure control, it is common for such publicly-available datasets to contain a measure of household incomes which is top-coded (i.e., right-censored) meaning that reported incomes above a certain threshold cannot be observed as measured and only an indicator of being above this income threshold is presented.

Deriving parametric distributions for the sample distribution of incomes y_i^{Obs} observed

under simple forms of measurement errors is the focus of early literature in the field. Models obtained from simple two-parameter distributions 'distorted' through classical measurement errors (i.e., independent of incomes) brought forward implications that would be in strong contrast with recent empirical observations: classical measurement errors can yield sample inequality estimates that overestimate inequality at the population level (e.g., see Krishnaji 1970, Hartley and Revankar 1974, Hinkley and Revankar 1977, Van Praag et al. 1983, Ransom and Cramer 1983, Chesher and Schluter 2002).

More recent literature, in change, has focused in characterizing under-reporting phenomena affecting income data. The robustly evidenced progresiveness of under-reporting with respect to income levels has resulted in more appropriate *nonclassical* parametric expressions for these measurement errors (e.g., see Gottschalk and Huynh 2010, Bourguignon 2018, Blanchet et al. 2022, and Flachaire et al. 2022) and has consistently found that high-income under-reporting yields sample inequality measures that underestimate inequality at the population level. This recent exploration of high-income under-reporting has given way to what are known as **replacing** corrections: incomes pressumed to be under-reported in the data are replaced by imputations from external data sources such as administrative tax data or by imputations from a model for the under-reporting mechanism. The 'corrected' data is then treated as a representative sample of the population's incomes following $f_y(.;\boldsymbol{\theta})$.

Another source of *MR* issues, that of missing observations, has also been treated under parametric approaches. The case of item non-response has received significantly more treatment than the more complex case of unit non-response (e.g., see Brunori et al. (2022) for a recent survey). The main aspect determining how to proceed concerns the distinction between observations Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR), following the works of Rubin (1976), Rubin (1977), and Greenlees et al. (1982). In the MCAR case the probability of a unit/item being missing in the data is independent of any characteristics of the unit and is constant across all units, inducing no particular biases to any distributional estimates from the data. The MAR case allows this probability to change with the characteristics of the unit but requires it to be independent of the unit's income level. Finally, the more complex MNAR case allows this probability to also change with the unit's income level and is therefore the only mechanism capable of representing the empirically evidenced negative relationship between response probabilities and income levels in survey data (e.g., Bollinger et al. 2019, Hlasny 2020).

The biases introduced by MAR or MNAR missing data mechanisms in distributional analysis have mostly been treated under the assumption that unit/item missingness is due to non-response. Namely, the assumption that conditional on being sampled high-income units are less likely to report their incomes (in the case of item non-response) or any information at all (in the case of unit non-response) than other units. This approach has motivated the use of **reweighting** corrections: the empirical distribution of incomes in the sample is reweighted by the related distribution of (imputed) response probabilities. Like with replacing, the reweighted data is then treated as a representative sample of the population's incomes following $f_y(.;\boldsymbol{\theta})$.

A particularly lacking aspect of the recent replacing/reweighting approaches in deal-

ing with *MR* is the lack of unified parametric frameworks integrating the modelling assumptions on the income distribution $f_y(.;\boldsymbol{\theta})$ and those on the under-reporting and/or non-response mechanisms. This has several consequences on the applicability and generalizability of these methods.

As a model for the data directly as it is observed, a unified parametric approach can allow for deriving expressions and estimation strategies suitable for microdata but also for other formats such as bracketed or grouped data (under known grouping mechanisms). Additionally, this may prove useful in dealing with the challenge that recent approaches in the literature face concerning the choice of correction quantities (i.e., the share of missing and/or under-reported incomes and their distribution). In general these quantities are hand-set by the analyst or are set to match quantities given from more reliable external data.

While setting correction quantities *ad hoc* relies entirely on the analyst's knowledge about the population's true income distribution, setting these quantities taking external data as reference poses several issues of its own. Firstly, it is not always the case that more reliable external data sources on incomes are available for research purposes as there may be access restrictions to such data or the data may suffer from *MR* issues of their own such as those induced by tax evasion and tax avoidance on administrative tax data. Secondly, even when external data is available it is generally the case that the population coverage and income components covered differ from those in the primary data available for the analysis and this implies that several compatibilizations must be made in transfering quantities from the former to complement the latter. This compatibilizations often come at the cost of forcing different income concepts to represent the same and of a loss in being able to quantify the statistical uncertainty around the resulting distributional estimates and in particular how these are affected by the inherent uncertainty concerning the correction quantities.

The parametric framework proposed in what follows builds on the recent literature exploring replacing and reweighting corrections for *MR* by integrating in a single distribution function both the population income distribution model $f_y(.;\boldsymbol{\theta})$ and any assumed form for measurement or missing data issues affecting incomes. Several formats of data and inference may be analyzed through the scope of this framework including that of learning about plausible values for the different *MR* correction quantities from the data itself and of integrating the uncertainty around these quantities into distributional estimates such as the Gini coefficient.

3 A parametric replacing and reweighting framework

Let individual i's *true* income be denoted by $y_i \sim f_{\boldsymbol{y}}(.;\boldsymbol{\theta})$, with probability density function (pdf) $f_{\boldsymbol{y}}(.;\boldsymbol{\theta})$ and cumulative distribution function (CDF) $F_{\boldsymbol{y}}(.;\boldsymbol{\theta})$, and consider a sample of *observed* individual incomes $\boldsymbol{y}^{Obs} = \{y_i^{Obs}\}_{i=1}^N$. This sample may be affected by two types of 'missing rich' issues: high-income under-reporting, in which case income y_i^{Obs} is observed but differs from y_i following under-reporting of high incomes, and non-response, in which case no income is observed for the individual i.

To introduce a parametric model for data under both of these possible issues, let

$\varphi(\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\nu})$ a **response probability function** defining the probability for an individual to report her income after being sampled from the population. In its most general formulation this probability, parametrized by the vector $\boldsymbol{\nu}$, may depend on the individual's income y_i and/or other characteristics $\boldsymbol{X}_{i,.}$ but also on others' incomes $\boldsymbol{y}$ and/or characteristics $\boldsymbol{X}$ more generally. Additionally, denote by $m(\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\eta})$ an **income reporting function**, defining the link between i's income y_i and her income reported in the data, if any, $y_i^{Obs} \equiv m(\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\eta})$. In particular, consider this latter function to be invertible, such that $y_i \equiv m^{-1}(\boldsymbol{y}^{Obs}, \boldsymbol{X}; \boldsymbol{\eta})$ is a **replacing function**.

Within this framework, we can relate i's income to her observed income y_i^{Obs}, if reported, and to some unobservable income y_i^{NObs}, in case of non-response, following[1]:

$$y_i = \begin{cases} m^{-1}(y_i^{Obs}, \boldsymbol{X}_i; \boldsymbol{\eta}) \ , \text{ with probability } \varphi(y_i, \boldsymbol{X}_i; \boldsymbol{\nu}) \\ y_i^{NObs} \ , \text{ with probability } 1 - \varphi(y_i, \boldsymbol{X}_i; \boldsymbol{\nu}) \end{cases}$$

If no measurement or non-response issues are believed to affect the data, then this amounts to setting $(m^{-1}(y_i^{Obs}, \boldsymbol{X}_i; \boldsymbol{\eta}), \varphi(y_i, \boldsymbol{X}_i; \boldsymbol{\nu})) \equiv (y_i^{Obs}, 1)$.

Whenever some form of measurement error is assumed to affect incomes in the data, then this may be introduced through a specific choice for the replacing function $m^{-1}(y_i^{Obs}, \boldsymbol{X}_i; \boldsymbol{\eta})$. This function serves the purpose of introducing any replacing or imputation step where i's income is set as a function of her observed income.

Any $m^{-1}(y_i^{Obs}, \boldsymbol{X}_i; \boldsymbol{\eta})$ representing progressive under-reporting of high incomes should imply an increasing and convex quantile ratio $r(i; \boldsymbol{\eta})$ defined as:

$$r(i; \boldsymbol{\eta}) = \frac{m^{-1}(y_{(i)}^{Obs}; \boldsymbol{\eta})}{y_{(i)}^{Obs}} \ , \ \frac{\partial r(i; \boldsymbol{\eta})}{\partial y_{(i)}^{Obs}} \geq 0 \ , \ \frac{\partial^2 r(i; \boldsymbol{\eta})}{\partial^2 y_{(i)}^{Obs}} \geq 0$$

with $y_{(i)}^{Obs}$ denoting the i-th quantile of $\boldsymbol{y}^{Obs}$. This restricts relative discrepancies between observed y_i^{Obs} and replaced $m^{-1}(y_i^{Obs}; \boldsymbol{\eta})$ incomes to be non-decreasing with incomes.

Recently popular replacing approaches can easily be expressed as deterministic forms for $m^{-1}(y_i^{Obs}, \boldsymbol{X}_i; \boldsymbol{\eta})$ including:

- **Piecewise linear quantile-ratio replacing (e.g., Flachaire et al. 2022):**

$$m^{-1}(y_{(i)}^{Obs}; \{\bar{p}_k\}_{k=1}^{K-1}, \{\beta_k\}_{k=1}^{K-1}, \{\delta_k\}_{k=1}^{K-1}) \equiv \begin{cases} y_{(i)}^{Obs} \ , \text{ if } p_{(i)} \leq \bar{p}_1 \\ y_{(i)}^{Obs} \times \sum_{k=1}^{K-1} \mathbf{1}(\bar{p}_k < p_{(i)} \leq \bar{p}_{k+1}) \times \underbrace{(\beta_k + \delta_k p_{(i)})}_{\substack{\text{Linear replacing weights} \\ \text{for incomes in} \\ \text{the } k\text{-th segment.}}} \end{cases}$$

  with $\delta_j \leq \delta_{j+1} < \infty \ , \ j = 1, ..., K-2$, $\bar{p}_K = 1$, and with $p_{(i)} = F_{\boldsymbol{y}}(y_{(i)}; \boldsymbol{\theta})$ denoting ordered-incomes individual $y_{(i)}^{Obs}$'s percentile in the population's income distribution[2]. In absence of missing data, the sample percentile $p_{(i)}^{Obs} \equiv \frac{(i)}{N} \ , \ (i) = 1, ..., N$ is

[1] For simplification reasons, all derivations in what follows are under the assumption that i's both response probabilities and reported income depend only on i's income and characteristics: $m^{-1}(\boldsymbol{y}^{Obs}, \boldsymbol{X}; \boldsymbol{\nu}) \equiv m^{-1}(y_i^{Obs}, \boldsymbol{X}_i; \boldsymbol{\nu})$, $\varphi(\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\nu}) \equiv \varphi(y_i, \boldsymbol{X}_i; \boldsymbol{\nu})$

[2] In what follows $\mathbf{1}(.)$ represents the identity function, taking value 1 whenever the condition it takes as argument holds true and 0 otherwise.

also a valid estimate of $F_{\boldsymbol{y}}(y_{(i)};\boldsymbol{\theta})$. The central assumption under this approach is that the quantile ratio $r(i;\boldsymbol{\eta}) = \frac{m^{-1}(y_{(i)}^{Obs};\{\bar{p}_k\}_{k=1}^{K-1},\{\beta_k\}_{k=1}^{K-1},\{\delta_k\}_{k=1}^{K-1})}{y_{(i)}^{Obs}}$ can be represented as a continuous piecewise linear function. This piecewise representation allows for progressive under-reporting of high incomes across segments $(\bar{p}_k;\bar{p}_{k+1}]$ of the income distribution at the cost of introducing 3 additional parameters $(\bar{p}_k;\beta_k;\delta_k)$ per segment.

- **Linear progressive under-reporting (LPU, Bourguignon 2018):**

$$m^{-1}(y_i^{Obs};\bar{p},\delta) \equiv y_i^{Obs}\times\left[1+\underbrace{\mathbf{1}(y_i^{Obs}>F_{\boldsymbol{y}}^{-1}(\bar{p};\boldsymbol{\theta}))}_{\substack{\text{Indiv. with observed incomes}\\ \text{above } \bar{p}\text{-th percentile under-report}}}\times\underbrace{\left(\frac{\delta(y_i^{Obs}-F_{\boldsymbol{y}}^{-1}(\bar{p};\boldsymbol{\theta}))}{1-\delta}\right)}_{\substack{\text{Under-reported amount linearly}\\ \text{increases with true incomes with slope } \delta}}\right]$$

with $\delta\in[0,1)$ and with $F_{\boldsymbol{y}}^{-1}(\bar{p};\boldsymbol{\theta})$ denoting the $\bar{p}$-th population income quantile. This replacing scheme assumes that all individuals with incomes above the $\bar{p}$-th percentile under-report their incomes in the observed sample and do so in a linearly progressive manner with under-reporting increasing by δ with every additional unit of income. The incomes quantile ratio implied under LPU $r(i;\boldsymbol{\eta}) = \frac{m^{-1}(y_{(i)}^{Obs};\bar{p},\delta)}{y_{(i)}^{Obs}}$ is strictly convex for income levels above $F_{\boldsymbol{y}}^{-1}(\bar{p};\boldsymbol{\theta})$.

- **Generalized Pareto replacing: (e.g., Atkinson and Piketty 2007, Chapter 2,, Jenkins 2017, Hlasny and Verme 2022, Charpentier and Flachaire 2022)**:

$$m^{-1}(y_i^{Obs};\mu,\sigma,\zeta) \equiv y_i^{Obs}\times\left[1+\underbrace{\mathbf{1}(y_i^{Obs}>\mu)}_{\substack{\text{Indiv. with observed incomes}\\ \text{above } \mu \text{ under-report}}}\times\underbrace{\left[\left(\frac{\left(1-\left(\frac{p_i-\bar{p}}{1-\bar{p}}\right)\right)^{-\zeta}-1}{\zeta}\right)\times\sigma-(y_i^{Obs}-\mu)\right]}_{\substack{\text{Observed incomes are}\\ \text{replaced by corresponding percentile}\\ \text{under a Generalized Pareto dist.}}}\right]$$

where (μ,σ,ζ) are respectively the location, scale, and shape parameters of a Generalized Pareto distribution $\mathcal{GPD}(\mu,\sigma,\zeta)$ with CDF given by:

$$F_{\boldsymbol{y}}(y_i;\mu,\sigma,\zeta) = \begin{cases} 1-\left(1+\frac{\zeta(y_i-\mu)}{\sigma}\right)^{-\frac{1}{\zeta}} & ,\ \text{if } \zeta\neq 0 \\ 1-e^{-\left(\frac{y_i-\mu}{\sigma}\right)} & ,\ \text{if } \zeta=0 \end{cases} \quad ,\ y_i>\mu$$

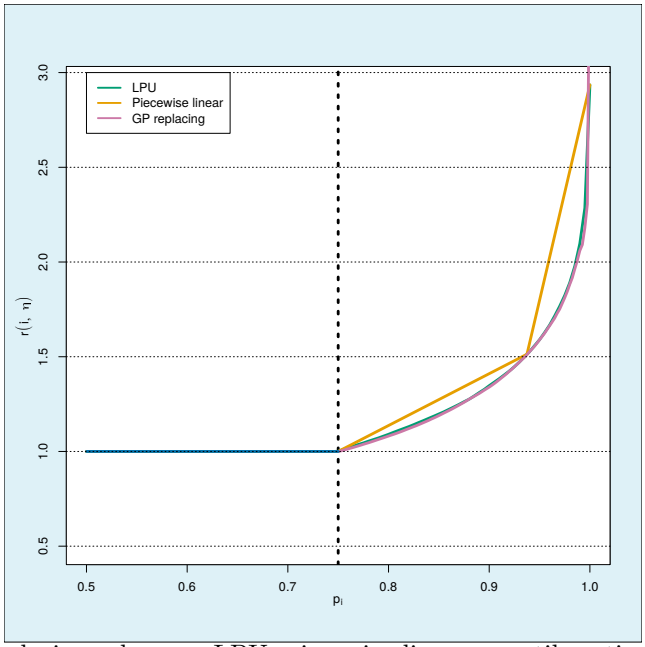
Under this replacing scheme, any income above a level μ is assumed to be under-reported. True incomes are assumed to follow a Generalized Pareto distribution (or some specific case such as the Pareto I (i.e., $\mathcal{GPD}(\frac{\sigma}{\zeta},\sigma,\zeta)\ ,\ \zeta>0$) or Pareto II (i.e., $\mathcal{GPD}(\mu,\sigma,\zeta)\ ,\ \zeta>0$)) above this income level. In absence of missing data issues an individual in the p_i-th sample percentile with $p_i>\bar{p}$ (equivalently, $y_i^{Obs}>\mu$) has true income corresponding to the $\left(\frac{p_i-\bar{p}}{1-\bar{p}}\right)$-th quantile on this Pareto

distribution. Similarly to LPU, the incomes quantile ratio implied under Generalized Pareto replacing $r(i;\boldsymbol{\eta}) = \frac{m^{-1}(y_{(i)}^{Obs};\mu,\sigma,\zeta)}{y_{(i)}^{Obs}}$ is strictly convex for income levels above μ, representing progressiveness of the under-reporting.

These common replacing schemes all exploit the assumption that under-reporting is a deterministic function of individual incomes (or their sample percentile/rank equivalently), and that individuals have the same rank in the population's income distribution as in the observed sample. It's also important to note that each specific replacing scheme implies within it specific assumptions on under-reporting behaviour at the individual level.

Figure 1 illustrates a comparative example of the quantile ratios $r(i,\boldsymbol{\eta})$ under these three common forms for $m^{-1}(.;\boldsymbol{\eta})$. The respective parameter values $\boldsymbol{\eta}$ are set to represent a same progressive under-reporting pattern: LPU affecting observed incomes from the .75-th percentile of the income distribution upwards. A first observation is that a piecewise linear approximation to this under-reporting pattern introducing six parameters in $\boldsymbol{\eta}$ in total (i.e., a linear approximation with two segments) is not flexible enough to correctly represent it. Secondly, replacing under a Generalized Pareto tail all incomes above the .75-th percentile can represent the reference LPU pattern accurately, except for the top of the income distribution. Finally, this similarity across GPD and LPU schemes suggests the latter as the more stable and parsimonious alternative of the two.

Figure 1: Quantile ratios under common replacing schemes



Note: Three common replacing schemes: LPU, piecewise linear quantile ratio, and Generalized Pareto replacing (GP in legend), as applied to a same income distribution following $y_i \sim GB2(2.257, 17393, 1, 1.033)$ (see following sections for details on the GB2 distribution) and affected by LPU with $\bar{p} = .75$ (represented by the dashed vertical line) and $\delta = .67$. The piecewise linear approximation was calibrated to fit this LPU pattern at the $\bar{p}_1 = .75$ and $\bar{p}_2 = .9375$ sample percentiles. The GPD coefficients ζ and σ were estimated conditional on μ being the .995-th sample quantile as a typical empirical practice (e.g., see Jenkins (2017)) and imposing finite variance ($\zeta < \frac{1}{2}$).

Further choices for $m^{-1}(y_i^{Obs}, \boldsymbol{X}_i; \boldsymbol{\eta})$ may exploit other individual characteristics $\boldsymbol{X}_i$ to represent larger heterogeneities in under-reporting patterns. A popular choice when data on individual consumption is available without reporting errors of its own is to define the income reporting function as an Engel curve (e.g., see Pissarides and Weber 1989, Lyssiotou et al. 2004, Hurst et al. 2014). Additionally, a stochastic component may be introduced in the definition of $m^{-1}(y_i^{Obs}, \boldsymbol{X}_i; \boldsymbol{\eta})$ to allow for heterogeneity in under-reporting behaviour across individuals with same level of incomes (e.g., Flachaire et al. 2022).

Concerning the modelling assumptions for the response probabilities $\varphi(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{\nu})$, the several possible types of missing data mechanisms may be considered, following Rubin (1976). If income non-response follows a random process which is unrelated to incomes $\boldsymbol{y}$ and other characteristics $\boldsymbol{X}$, then the mechanism corresponds to a MCAR process. In the MCAR case, we observe incomes $\boldsymbol{y}^{Obs} = \{y_i^{Obs}\}_{i=1}^N$ which are a random sample of the population's incomes and therefore no particular bias is induced by the missing data. A simple MCAR mechanism is such that $\varphi(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{\nu}) \equiv \varphi(\boldsymbol{y}_i, \boldsymbol{X}_i; p) \equiv p \ , \ p \in (0, 1]$, where all individuals are just as likely to report incomes after they have been sampled.

A second potential mechanism concerns the case where non-response in incomes are not random but where the misingness can be fully explained by other non-missing characteristics of the individuals and/or by the observed incomes, i.e., $\varphi(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{\nu}) \equiv \varphi(\boldsymbol{y}_i^{Obs}, \boldsymbol{X}_i; \boldsymbol{\nu})$. This mechanism represents an MAR process and is an appropriate representation for scenarios of item non-response, where sampled individuals report information about their characteristics $\boldsymbol{X}_i$ but not about their income, as long as their unobserved income y_i^{Nobs} is unnecessary to account for the non-random non-response probabilities. MAR mechanisms are usually dealt with in analysis through multiple imputations of incomes for those individuals in the data with missing incomes but observed characteristics.

Finally, it may be the case that response probabilities may not be fully accounted for from observed information. For instance, it may be the case that the reason why individuals do not report their incomes in the data has everything to do with their unobserved level of incomes y_i^{NObs}. This corresponds to the MNAR scenario and is particularly complex to deal with, as it may include non-random unit non-response mechanisms, where sampled individuals do not report neither incomes nor characteristics and where their unobserved incomes y_i^{NObs} are a determinant of this.

Forms for $\varphi(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{\eta})$ suitable for MAR mechanisms have been the focus of the recent survey in Brunori et al. (2022). Recently popular reweighting approaches allowing for dealing also with MNAR mechanisms can easily be expressed as deterministic forms for $\varphi(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{\eta})$ including:

- **Right-truncation (e.g., Alvaredo 2011, Jorda and Niño-Zarazúa 2019):**

$$\varphi(y_{(i)}; t, \alpha) \equiv \begin{cases} \alpha \ , \ \text{if } p_{(i)} \leq t \\ 0 \ , \ \text{if } p_{(i)} > t \end{cases}$$

  which amounts to assuming that any and all individuals above the t-th percentile on the population income distribution will not report incomes in the data, while anyone below this threshold will report an income with probability α. The limiting

case $\alpha \to 1$ corresponds to assuming that any unit with income below the t-th percentiles will always report an income when sampled.

- **Regional non-response reweighting (e.g., Korinek et al. 2007, Hlasny and Verme 2018)**:

$$\varphi(y_i, \boldsymbol{X}_i; \boldsymbol{\beta}) \equiv \frac{e^{g(y_i, \boldsymbol{X}_i; \boldsymbol{\beta})}}{1 + e^{g(y_i, \boldsymbol{X}_i; \boldsymbol{\beta})}}$$

with $g(y_i, \boldsymbol{X}_i; \boldsymbol{\beta})$ being a twice continuously differentiable function of observed unit i's characteristics parametrized by the vector $\boldsymbol{\beta}$. The comparative analysis in Hlasny and Verme (2015) suggests a simple logarithmic form for g taking as input a linear combination of income y_i and region indicator variables to be equally efficient as more complex specifications in many scenarios. This approach infers response probabilities for units from modelling the relationship between non-response rates and units' characteristics at aggregate (i.e., regional) levels, when this information is available. The key assumption is that individual characteristics relate to individual response probabilities in the same way that they do at the aggregate level (i.e., that ecological inference is feasible), such that individual response probabilities $\varphi(y_i, \boldsymbol{X}_i; \boldsymbol{\beta})$ may be properly estimated and used for the purpose of reweighting the observed data.

- **Income-proportional reweighting (e.g., Blanchet et al. 2022):**

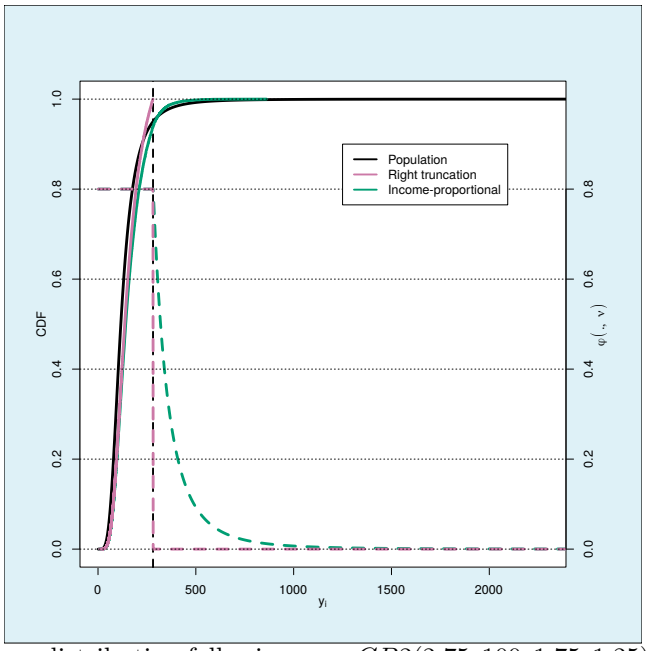
$$\varphi(y_i; \gamma_0, \gamma_1, t, \alpha) = \begin{cases} e^{\gamma_0} (y_i)^{-\gamma_1} \ , \ \gamma_1 > 0 \ , \ \text{if } p_{(i)} > t \\ \alpha \ , \ \text{if } p_{(i)} \leq t \end{cases}$$

This scheme corresponds to a non-response mechanism where individuals with true incomes above the t-th percentile have increasingly lower response probabilities, with the parameter γ_1 representing the income elasticity of non-response (i.e., how much response probabilities decrease with an increase in incomes of 1%). For a given value of such elasticity, γ_0 serves as an intercept to assure the continuity of $\varphi(y_i; \gamma_0, \gamma_1, t, \alpha)$ at t. Similarly to right-truncation, the parameter α represents the (constant) response probability for units with incomes below the t-th percentile. Moreover, this reweighting scheme includes the right-truncation $\varphi(y_{(i)}; t, \alpha)$ as the limiting case $\gamma_1 \to \infty$. Figure 2 provides an illustratory example of how these two cases relate and their resulting contrasts with the population's income distribution.

Both replacing and reweighting corrections may interact within this framework in a straightforward manner. Response probabilities are modelled as a function of true incomes $\boldsymbol{y}$, yet these may differ from observed incomes $\boldsymbol{y}^{Obs}$ following an assumed form for $m^{-1}(y_i^{Obs}, \boldsymbol{X}_i; \boldsymbol{\eta})$. The interaction allowing for modelling MNAR non-response mechanisms through observed incomes directly simply amounts to the composite function $\varphi(m^{-1}(y_i^{Obs}, \boldsymbol{X}_i; \boldsymbol{\eta}), \boldsymbol{X}_i; \boldsymbol{\nu})$.

A crucial question this parametric framework allows to answer is: given i) an assumed form for the population income distribution $f_{\boldsymbol{y}}(.; \boldsymbol{\theta})$, ii) an assumed income reporting form $m(y_i, \boldsymbol{X}_i; \boldsymbol{\eta})$, and iii) an assumed response probability function $\varphi(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{\nu})$, then what distribution $f_{\boldsymbol{y}^{Obs}}(y_i^{Obs}; \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$ will observed incomes under this framework follow?

Figure 2: Income-proportional reweighting schemes



Note: A population income distribution following $y_i \sim GB2(2.75, 100, 1.75, 1.25)$ and its corresponding CDF under two cases of income-proportional non-response schemes: Right truncation, with parameter values set at $(\alpha, t) = (.8, .95)$, and income-proportional with parameter values $(\gamma_1, \alpha, t) = (3.75, .8, .95)$ which requires $\gamma_0 = 20.92$ for continuity. Solid lines represent respective CDFs, on the left axis, and dashed lines represent response probabilities, on the right axis. The dashed vertical line represents the t-th population percentile.

This distribution can be obtained applying the deterministic transformation $m^{-1}(.;\boldsymbol{\eta})$ to $f_{\boldsymbol{y}}(.;\boldsymbol{\theta})$ and reweighting the resulting density by the response probabilities $\varphi(.;\boldsymbol{\nu})$, yielding the relationship[3]:

$$f_{\boldsymbol{y}^{Obs}}(y_i^{Obs};\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\nu}) = \frac{\overbrace{f_{\boldsymbol{y}}(m^{-1}(y_i^{Obs};\boldsymbol{\eta});\boldsymbol{\theta}) \times \left(\frac{\partial m^{-1}(y_i^{Obs};\boldsymbol{\eta})}{\partial y_i^{Obs}}\right)}^{\text{Reporting function: Replacing transformation of } \boldsymbol{y}} \times \overbrace{\varphi(m^{-1}(y_i^{Obs};\boldsymbol{\eta});\boldsymbol{\nu})}^{\text{Non-response: Reweighting of } f_{\boldsymbol{y}}}}{\underbrace{\int f_{\boldsymbol{y}}(m^{-1}(y_i^{Obs};\boldsymbol{\eta});\boldsymbol{\theta}) \times \varphi(m^{-1}(y_i^{Obs};\boldsymbol{\eta});\boldsymbol{\nu}) \times \left(\frac{\partial m^{-1}(y_i^{Obs};\boldsymbol{\eta})}{\partial y_i^{Obs}}\right) dy^{Obs}}_{\text{Normalizing constant}}} \quad (1)$$

The main application of the result in (1) is that of parametrically integrating all assumptions about the population income distribution and the 'missing rich' issues affecting the data in a model for the observed data itself. This model for the data constitutes therefore a standard case of continuous model expansion to accomodate for non-response or measurement errors (e.g., see Nandram and Choi 2002, Gustafson 2005, Gelman et al. 2013, Chapter 7,). Fitting such a model to data is an attempt at identifying separately features characteristic of the population income distribution, captured by the $\boldsymbol{\theta}$ vector, and features representing the 'missing rich' forms pressumed to affect the data, captured by the $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$ vectors.

There are several virtues to integrating the replacing and/or reweighting corrections

[3]For simplicity and without loss of generality, only forms of the type $m(.;\boldsymbol{\eta}) \equiv m(y_i;\boldsymbol{\eta})$ and $\varphi(.;\boldsymbol{\eta}) \equiv \varphi(y_i;\boldsymbol{\eta})$ are considered in what follows.

considered relevant into a model to be taken to the data *as-is*. Firstly, because the correction quantities are completely defined through $\boldsymbol{\eta}$ and/or $\boldsymbol{\nu}$ this approach guarantees that all corrections are done on the income concept and population being analyzed. This avoids the issue of manipulating these concepts to be compatible with correction quantities defined in terms of different income concepts or population. On a related note, if external data informative on the forms and magnitudes of 'missing rich' are available then these should be introduced by specifying adequate representantions $m^{-1}(y_i^{Obs};\boldsymbol{\eta})$ and $\varphi(y_i^{Obs};\boldsymbol{\nu})$ and setting $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$ to quantify these magnitudes.

A second virtue of this integrated approach is that it makes it feasible to compute measures of uncertainty, such as standard errors, for the estimated parameter values. Importantly, these measures of uncertainty can allow for probabilistic assessment of the relevance of any corrections considered, as well as allowing for considering the uncertainty in the estimated features of the population's income distribution that also considers the uncertainty surrounding the corrections quantities.

Thirdly, the 'building blocks' nature of the framework allows for exploring several candidate forms for replacing and/or reweighting corrections leaving other components unchanged in a straightforward manner. In particular, this allows for studying the robustness of the estimated $\boldsymbol{\theta}$ to different assumptions on the form of 'missing rich' affecting the data.

Finally, stating the model as a properly defined parametric distribution implies that all observed units are re-weighted under any assumed form for $\phi(.;\boldsymbol{\nu})$, either directly through the reweighting of units prone to non-response in the numerator of (1) (i.e., through downweighting the density at their respective level of income with respect to the population density.) or indirectly through the correction for missing observations in the normalization constant of (1). This 'indirect' reweighting accomodates for the fact that if some units are under-represented in the data due to higher non-response probabilities then necessarily the rest of units are over-represented and therefore need to be reweighted under any correction for these non-response probabilities.

Making inference about the features of the population income distribution and the 'missing rich' aspects of the data simultaneously poses several challenges. Issues of identifiability, in particular, require attention as a given model specified following (1) might fit equally well a sample of observed incomes for very different values of the $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\nu}$ parameters, making inference on them invalid. The type of continuous model expansion underlying (1) to introduce uncertainty about the specific form and magnitudes of the 'missing rich' issues affecting the data falls in line with previous empirical strategies within Bayesian inference (e.g., see Nandram and Choi 2002). The use of prior probabilities on parameter values under a Bayesian approach can overcome some identifiability issues. The following section details a Bayesian inference approach for this purpose which can exploit external information on the 'missing rich' correction quantities in dealing with this.

4 Parameter inference under 'missing rich'

Under the framework developed in the previous section, inference on the population's income distribution $f_{\boldsymbol{y}}(.;\boldsymbol{\theta})$ is made through inference on the values of the $\boldsymbol{\theta}$ vector given the sample of observed incomes $\boldsymbol{y}^{Obs}$. This task is considerably less complex whenever the correction quantities $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$ are given fixed values. However, it is rarely the case that sound candidate values for these quantities are available. The central challenge in making inference on $\boldsymbol{\theta}$ is therefore to exploit the framework under an empirical strategy that can properly estimate these parameters but also $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$ at the same time.

The goal is to learn about which values for the parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\theta}} \subseteq \mathbb{R}^{dim(\boldsymbol{\theta})}$, $\boldsymbol{\eta} \in \boldsymbol{\Theta}_{\boldsymbol{\eta}} \subseteq \mathbb{R}^{dim(\boldsymbol{\eta})}$, and $\boldsymbol{\nu} \in \boldsymbol{\Theta}_{\boldsymbol{\nu}} \subseteq \mathbb{R}^{dim(\boldsymbol{\nu})}$ are more likely to have generated the observed data $\boldsymbol{y}^{Obs}$ than others within some region of possible values $\boldsymbol{\Theta} \equiv \boldsymbol{\Theta}_{\boldsymbol{\theta}} \times \boldsymbol{\Theta}_{\boldsymbol{\eta}} \times \boldsymbol{\Theta}_{\boldsymbol{\nu}}$.

In the Bayesian framework, this information takes the form of a posterior probability distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$ defined by two main components under Bayes' theorem. Firstly, all prior beliefs about the values of the $(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$ parameters must be ellicited through a prior probability distribution $p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$ over $\boldsymbol{\Theta}$. Secondly, for any fixed value for the parameters $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}})$ the model's likelihood $\mathcal{L}(\boldsymbol{y}^{Obs}|\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}})$ quantifies how likely the observed data $\boldsymbol{y}^{Obs}$ is to have been generated from $f_{\boldsymbol{y}^{Obs}}(.;\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}})$[4]. $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$ is then a probability distribution proportional to the prior probability distribution *updated* (or reweighted, equivalently) by the likelihood function:

$$\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs}) \propto \mathcal{L}(\boldsymbol{y}^{Obs}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}) \times p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}) \tag{2}$$

As an evidence-weighted conversion of prior beliefs, the information contained in the $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$ posterior distribution can be interpreted as all remaining uncertainty on the values of $(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$ after having 'learnt' from the data through the likelihood $\mathcal{L}(\boldsymbol{y}^{Obs}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$. Whenever the data are informative about these parameters, the posterior distribution reflects less uncertainty around their values than that in $p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$.

Estimating a posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$ for the model parameters faces several complexities. As is usual in most Bayesian inference settings, it is rarely the case that $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$ admits a known form given a model $\mathcal{L}(\boldsymbol{y}^{Obs}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$ and a prior $p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$. This is typically circumvented by studying the posterior distribution through samples generated to converge to $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$ under the Monte Carlo principle[5] or the Markov Chain Monte Carlo (MCMC) extension of this principle (e.g., see Gelman et al. 2013, Chapter 11).

A second complexity in estimating $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$ concerns the possible 'non-identifiability' of at least some of the parameters in $(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$. As an illustratory example of this issue, consider a model specified following (1) with a parameter $\lambda_{\boldsymbol{\theta}} \in \boldsymbol{\theta}$ ruling the right tail of the $f_{\boldsymbol{y}}(.;\boldsymbol{\theta})$ income distribution and a replacing correction $m^{-1}(.;\boldsymbol{\eta})$ with parameter $\lambda_{\boldsymbol{\eta}} \in \boldsymbol{\eta}$ also affecting only the right tail. It can be the case that a same

[4]For example, in the case of $\boldsymbol{y}^{Obs} = \{y_i^{Obs}\}_{i=1}^N$ being N independent observations their joint likelihood follows $\mathcal{L}(\boldsymbol{y}^{Obs}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}) = \prod_{i=1}^N f_{\boldsymbol{y}^{Obs}}(y_i^{Obs};\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$

[5]The Monte Carlo principle states that any quantity of $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$ which can be expressed as an expectation can be studied through a sufficiently large sample of J independent draws $\{(\tilde{\boldsymbol{\theta}}^{(j)}, \tilde{\boldsymbol{\eta}}^{(j)}, \tilde{\boldsymbol{\nu}}^{(j)})\}_{j=1}^J$ from this distribution $(\tilde{\boldsymbol{\theta}}^{(j)}, \tilde{\boldsymbol{\eta}}^{(j)}, \tilde{\boldsymbol{\nu}}^{(j)}) \sim \pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$

sample of incomes $\boldsymbol{y}^{Obs}$ may be equally well fit under two different parameter values $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}}) \in \boldsymbol{\Theta}$ and $(\tilde{\boldsymbol{\theta}}', \tilde{\boldsymbol{\eta}}', \tilde{\boldsymbol{\nu}}') \in \boldsymbol{\Theta}$ including $(\tilde{\lambda}_{\boldsymbol{\theta}}, \tilde{\lambda}_{\boldsymbol{\eta}})$ and $(\tilde{\lambda}'_{\boldsymbol{\theta}}, \tilde{\lambda}'_{\boldsymbol{\eta}})$ respectively. This can render the model incapable of separately identifying variations in high incomes in $\boldsymbol{y}^{Obs}$ that would occur with changes in $\lambda_{\boldsymbol{\theta}}$ and those due to $\lambda_{\boldsymbol{\eta}}$.

If $\boldsymbol{y}^{Obs}$ is not informative about differences in the respective likelihoods $\mathcal{L}(\boldsymbol{y}^{Obs}|\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}})$ and $\mathcal{L}(\boldsymbol{y}^{Obs}|\tilde{\boldsymbol{\theta}}', \tilde{\boldsymbol{\eta}}', \tilde{\boldsymbol{\nu}}')$, then prior beliefs on these values will not be updated. The respective posterior probabilities $\pi(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}}|\boldsymbol{y}^{Obs})$ and $\pi(\tilde{\boldsymbol{\theta}}', \tilde{\boldsymbol{\eta}}', \tilde{\boldsymbol{\nu}}'|\boldsymbol{y}^{Obs})$ will therefore be dominated entirely by differences in prior beliefs $p(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}})$ and $p(\tilde{\boldsymbol{\theta}}', \tilde{\boldsymbol{\eta}}', \tilde{\boldsymbol{\nu}}')$. If available, external information about the plausibility of $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}})$ and $(\tilde{\boldsymbol{\theta}}', \tilde{\boldsymbol{\eta}}', \tilde{\boldsymbol{\nu}}')$ may be exploited to set an informative prior distribution $p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$ giving a lower prior probability to the one set of parameter values less compatible with this external data amongst the two. Informative priors are a way of exploiting prior knowledge to justify differences in posterior densities for parameter values where $\boldsymbol{y}^{Obs}$ is uninformative through the model $\mathcal{L}(\boldsymbol{y}^{Obs}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$.

Returning to the illustratory example above, consider an application of (1) as a model for a survey's sample on incomes $\boldsymbol{y}^{Obs} = \{y_i^{Obs}\}_{i=1}^N$ integrating a GB2 income distribution $f_{\boldsymbol{y}}^{GB2}(.;\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\alpha, \beta, p, q)$ with an LPU form for $m^{-1}(.;\boldsymbol{\eta})$, $\boldsymbol{\eta} = (\bar{p}, \delta)$. LPU affects only the tail above the $\bar{p}$-th percentile of the income distribution, while the p and q parameters of the GB2 distribution rule its right tail. This allows for identifiability issues as described above, as there might be configurations 'trading' values of p and q with values of $\bar{p}$ and δ while representing two observably identical income distributions. In this example, external information might be introduced in the form of prior probabilities by setting the marginal prior distributions for $\bar{p}$ and δ around previous empirical findings on 'missing rich' issues in similar settings[6]

Several sampling algorithms can be devised to obtain samples $\{(\tilde{\boldsymbol{\theta}}^{(j)}, \tilde{\boldsymbol{\eta}}^{(j)}, \tilde{\boldsymbol{\nu}}^{(j)})\}_{j=1}^J$ from $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$ under a model following (1) and an informative prior $p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$. The Metropolis-Hastings (MH) algorithm defines a type of MCMC sampler suitable for estimating parametric income distribution models in several contexts (e.g., see Chotikapanich and Griffiths 2000, Peters and Sisson 2006, Chotikapanich and Griffiths 2008).

A standard MH sampler for the joint parameter vector $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$ is possible following algorithm (1) below. Such an MH algorithm yields as output a sample $\{\tilde{\boldsymbol{\theta}}^{(j)}, \tilde{\boldsymbol{\eta}}^{(j)}, \tilde{\boldsymbol{\nu}}^{(j)}\}_{j=1}^J$ resulting from a global exploration of the support of $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$ through local accept-reject steps. Any j-th, $j = 1, ..., J$, local accept-reject step is defined by the MH acceptance probability:

$$\rho^{(j)} = \min\left\{1, \frac{\pi(\tilde{\boldsymbol{\phi}}^{(j)}|\boldsymbol{y}^{Obs}) \times g(\tilde{\boldsymbol{\phi}}^{(j-1)}, \tilde{\boldsymbol{\phi}}^{(j)})}{\pi(\tilde{\boldsymbol{\phi}}^{(j-1)}|\boldsymbol{y}^{Obs}) \times g(\tilde{\boldsymbol{\phi}}^{(j)}, \tilde{\boldsymbol{\phi}}^{(j-1)})}\right\} \ , \ \boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$$

[6]For example, in their study comparing household survey incomes to linked tax return data for Uruguay Flachaire et al. (2022) find evidence of progressive under-reporting potentially affecting the survey data above $\bar{p} = .50$. In studying similar linked data for the Austrian case, Angel et al. (2019) find evidence of progressive under-reporting of wages potentially affecting their survey above the $\bar{p} = .50$ percentile. The degree of progresiveness of under-reporting can be quantified in terms of δ under a linear approximation to the observed under-reporting patterns.

with $g(\tilde{\boldsymbol{\phi}}^{(j)}, \tilde{\boldsymbol{\phi}}^{(j-1)})$ denoting a candidate function from which the j-th candidate value $\tilde{\boldsymbol{\phi}}^{(j)}$ is sampled, given the previously retained value $\tilde{\boldsymbol{\phi}}^{(j-1)}$.

Algorithm 1: A Metropolis-Hastings algorithm (***MH***).

Initialization:
Until $\mathcal{L}(\boldsymbol{y}^{Obs}|\boldsymbol{\phi}^{(0)}) > 0$:
  Sample $\tilde{\boldsymbol{\phi}}^{(0)}$ from $p(\boldsymbol{\phi})$
Sampling: **for** $j = 1, ..., J$ **do**
Sample $\tilde{\boldsymbol{\phi}}^{(j)} \sim g(\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}^{(j-1)})$ from the candidate g
Accept and store $\tilde{\boldsymbol{\phi}}^{(j)}$ with probability:

$$\rho^{(j)} = \min \left\{ 1, \frac{\overbrace{\mathcal{L}(\boldsymbol{y}^{Obs}|\tilde{\boldsymbol{\phi}}^{(j)}) \times p(\tilde{\boldsymbol{\phi}}^{(j)})}^{\propto \pi(\tilde{\boldsymbol{\phi}}^{(j)}|\boldsymbol{y}^{Obs}) \text{ under (2)}} \times g(\tilde{\boldsymbol{\phi}}^{(j-1)}, \tilde{\boldsymbol{\phi}}^{(j)})}{\mathcal{L}(\boldsymbol{y}^{Obs}|\tilde{\boldsymbol{\phi}}^{(j-1)}) \times p(\tilde{\boldsymbol{\phi}}^{(j-1)}) \times g(\tilde{\boldsymbol{\phi}}^{(j)}, \tilde{\boldsymbol{\phi}}^{(j-1)})} \right\}$$

▷ e.g., if $u^{(j)} \leq \rho^{(j)}$ where $u^{(j)}$ is a draw from a $Uniform(0,1)$ distribution
otherwise store $\tilde{\boldsymbol{\phi}}^{(j)} = \tilde{\boldsymbol{\phi}}^{(j-1)}$
end

A common choice of candidate function is that of the Adaptive Random-Walk Metropolis (***AM***) algorithm (Haario et al., 2001). In this case the proposal $g \equiv g_{\Sigma}$ is defined by the following adaptive random walk process:

$$(\boldsymbol{\theta}^{(j)}, \boldsymbol{\eta}^{(j)}, \boldsymbol{\nu}^{(j)}) \equiv \boldsymbol{\phi}^{(j)} \sim g_{\Sigma^{(j-1)}}(\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}^{(j-1)}) \Rightarrow \tilde{\boldsymbol{\phi}}^{(j)} = \tilde{\boldsymbol{\phi}}^{(j-1)} + \tilde{\boldsymbol{\epsilon}}^{(j)}$$
$$\boldsymbol{\epsilon}^{(j)} \sim N_d(0, \Sigma^{(j-1)})$$
$$\Sigma^{(j-1)} = \begin{cases} \Sigma^{(0)} \ , \text{ if } j \leq J_0 \\ s_d \times \frac{1}{(j-1)} \left(\sum_{i=1}^{(j-1)} \tilde{\boldsymbol{\phi}}^{(i)} \tilde{\boldsymbol{\phi}}^{(i)'} - i \times \bar{\boldsymbol{\phi}} \bar{\boldsymbol{\phi}}' \right) + s_d \times \chi \times I_d \ , \text{ if } j > J_0 \ , \ 0 < \chi \ll 1 \end{cases}$$

with $\bar{\boldsymbol{\phi}}$ denoting the mean value of all draws up to and including the $(j-1)$-th and with s_d suggested, following Gelman et al. (1996), to be set to $s_d = \frac{2.4^2}{d}$ where d is the number of parameters in $\boldsymbol{\phi}$[7].

Under this proposal distribution the j-th candidate value $\tilde{\boldsymbol{\phi}}^{(j)}$ is obtained by sampling from a multivariate Gaussian distribution centered at the previously retained draw $\tilde{\boldsymbol{\phi}}^{(j-1)}$ and with covariance matrix $\Sigma^{(j-1)}$. Being initially set to a given matrix $\Sigma^{(0)}$, this covariance matrix starts adapting exploiting all past draws after a sufficiently large initial period J_0 following the sample covariance matrix. An (***AM***) algorithm can thus focus on sampling more densely in regions near values $\tilde{\boldsymbol{\phi}}$ with high posterior density and less

[7]The addition of the diagonal matrix $\chi \times I_d$ is needed with an insignificantly small but non-zero χ to assure the non-singularity of $\Sigma^{(j-1)}$ and assure the convergence of the MCMC sampling distribution to $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$.

densely in regions of low posterior density. It is also possible to extend the scope of the local accept-reject exploration by sampling M candidates at once from $g_{\Sigma^{(j-1)}}(\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}^{(j-1)})$ in the spirit of the multiple-try Metropolis sampler of Liu et al. (2000).

Implementing the (***AM***) algorithm requires being able to compute the likelihood function $\mathcal{L}(\boldsymbol{y}^{Obs}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$. For a model following (1), the joint likelihood for a sample of independent microdata $\boldsymbol{y}^{Obs} = \{y_i^{Obs}\}_{i=1}^N$ can be computed as $\mathcal{L}(\boldsymbol{y}^{Obs}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}) = \prod_{i=1}^N f_{\boldsymbol{y}^{Obs}}(y_i^{Obs}; \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$. However, a computable likelihood function may not be available in several contexts, such as when data on incomes is only available at group level (e.g., see Kobayashi and Kakamu 2019, Eckernkemper and Gribisch 2021). A more general MCMC sampling algorithm, requiring only being able to simulate data from a model following (1) is available through the Approximate Bayesian Computation (ABC) approach (e.g., see Kobayashi and Kakamu 2019, Silva 2023).

ABC constitutes a class of simulation-based Bayesian inference methods which approximate the unavailable likelihood $\mathcal{L}(\boldsymbol{y}^{Obs}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$ through simulating data. This approximation requires a way to assess how closely the observed data $\boldsymbol{y}^{Obs}$ resembles data simulated from the model $\tilde{\boldsymbol{y}}^{Obs} \sim f_{\boldsymbol{y}^{Obs}}(.; \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}})$ for any given parameter values $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}}) \in \boldsymbol{\Theta}$[8]. Following Silva (2023), this comparison may be done by representing $\boldsymbol{y}^{Obs}$ and $\tilde{\boldsymbol{y}}^{Obs}$ through their respective empirical Generalized Lorenz curves (GLC, Shorrocks 1983) defined as[9]

$$GLC_k^{Obs} = \underbrace{\frac{\sum_{i=1}^k y_{(i)}^{Obs}}{\sum_{i=1}^N y_{(i)}^{Obs}}}_{s_k^{Obs}} \times \underbrace{\frac{1}{N}\sum_{i=1}^N y_{(i)}^{Obs}}_{\mu^{Obs}} = \frac{\sum_{i=1}^k y_{(i)}^{Obs}}{N} \ , \ k = 1, ..., N \ , \ GLC_0^{Obs} = 0$$

with s_k^{Obs} denoting the cumulative income share up to the k-th observation in the ordered sample and μ^{Obs} denoting the sample average income.

Given the observed-data GLC, denoted $\{GLC_k^{Obs}\}_{k=1}^N$, and the analogue GLC from a simulated sample, denoted $\{GLC_k^{Obs}(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}})\}_{k=1}^N$, the overall degree of discrepancy between both empirical income distributions may be summarized by the following unidimensional metric:

$$\begin{aligned} d(\boldsymbol{y}^{Obs}, \tilde{\boldsymbol{y}}^{Obs}) &= d(\{GLC_k^{Obs}\}_{k=1}^N, \{GLC_k^{Obs}(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}})\}_{k=1}^N) \\ &= \sum_{k=1}^N |(GLC_k^{Obs} - GLC_{k-1}^{Obs}) - (GLC_k^{Obs}(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}}) - GLC_{k-1}^{Obs}(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}})| \end{aligned}$$

which corresponds to the empirical Wasserstein-1 distance (Kantorovich, 1939) in the case of microdata[10]. Explored in the context of ABC by Bernton et al. (2019), this distance summarizes the absolute discrepancies between all order statistics across observed and simulated data $|y_{(i)}^{Obs} - \tilde{y}_{(i)}^{Obs}| \ , \ i = 1, ..., N$.

In approximating $\mathcal{L}(\boldsymbol{y}^{Obs}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$ for an ABC implementation of the (***AM***) algorithm, parameter values $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}}) \in \boldsymbol{\Theta}$ yielding simulated data $\tilde{\boldsymbol{y}}^{Obs}$ resembling $\boldsymbol{y}^{Obs}$ more closely

[8]For simplicity, it is assumed in what follows that simulated data are in the form of independent microdata $\tilde{\boldsymbol{y}} = \{\tilde{y}_i^{Obs}\}_{i=1}^N$.

[9]For a grouped-data formulation see Silva (2023).

[10]See the derivations in Appendix A for a grouped-data implementation of this discrepancy.

under $d(.,.)$ than others should be given larger importance. This is commonly introduced exploiting a kernel function K_τ giving increasingly larger weight to parameter values with a lower discrepancy $\varepsilon \equiv d(\boldsymbol{y}^{Obs}, \tilde{\boldsymbol{y}}^{Obs})$. A common 'smooth' kernel for this purpose is the Gaussian kernel (e.g., see Ratmann 2010):

$$K_\tau^{gauss}(\varepsilon) = \frac{1}{\tau} \times \frac{1}{\sqrt{2\pi}} \times \exp\left\{-\frac{1}{2}\left(\frac{\varepsilon}{\tau}\right)^2\right\} \ , \ \varepsilon \equiv d(\boldsymbol{y}^{Obs}, \tilde{\boldsymbol{y}}^{Obs})$$

. Under this kernel the ABC discrepancies are weighted following a Normal distribution centered at zero (i.e., highest weight is given to values $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\nu}}) \in \boldsymbol{\Theta}$ exactly reproducing $\boldsymbol{y}^{Obs}$), and with a standard deviation of τ.

The bandwidth parameter τ rules the strictness of the ABC approximation to $\mathcal{L}(\boldsymbol{y}^{Obs}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$ by defining how the weights $K_\tau^{gauss}(\varepsilon)$ decrease with an increase in the discrepancy ε. The approximation is exact when $\tau \to 0$, as only parameter values which exactly reproduce the observed income distribution are given a non-zero weight, and $\tau \to \infty$ amounts to considering any and all parameter values in $\boldsymbol{\Theta}$ equally likely to have generated the observed data (i.e., the likelihood is approximated as a flat function).

Following Silva (2023), an ABC (***AM***) algorithm with these settings can be devised extending the Marjoram et al. (2003) ABC implementation of the (***MH***) algorithm. Algorithm (2) below presents a possible implementation, denoted (***ABC-AM***) in what follows.

At any j-th step, the (***ABC-AM***) algorithm draws M candidate parameter values $\{\tilde{\boldsymbol{\phi}}^{(m)}\}_{m=1}^M$ from the adaptive proposal $g_{\Sigma^{(j-1)}}$, simulates a single income distribution from the model for each such candidate, and computes their respective discrepancies with respect to the observed income distribution. The candidate with the lowest discrepancy is then taken as the j-th candidate $\tilde{\boldsymbol{\phi}}^{(j)}$, along with its associated ABC discrepancy $\tilde{\varepsilon}^{(j)}$, in the same spirit as Clarté et al. (2021). Finally, the MH accept-reject rule is computed with respect to the ABC approximation $K_\tau(\tilde{\varepsilon}^{(j)})$ of the likelihood.

Introducing the ABC approximation $K_\tau(\tilde{\varepsilon}^{(j)})$ to the likelihood $\mathcal{L}(\boldsymbol{y}^{Obs}|\tilde{\boldsymbol{\theta}}^{(j)}, \tilde{\boldsymbol{\eta}}^{(j)}, \tilde{\boldsymbol{\nu}}^{(j)})$ implies sampling from an ABC approximation to the target posterior distribution $\pi_\tau(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\{GLC_k^{Obs}\}_{k=1}^N)$. This posterior distribution might differ from that in (2) for several reasons. One first source of differences lies on the quality of the approximation. The main determinant of this is the choice for the bandwidth parameter τ. In practice, the choice for this bandwidth results from calibrating the sampling algorithm through several initial runs balancing strictness of the approximation and computational cost.

The second main source for differences between $\pi_\tau(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\{GLC_k^{Obs}\}_{k=1}^N)$ and $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{y}^{Obs})$ concerns the possible loss of information due to summarizing the data through the GLC and not through the microdata directly. If what can be learnt about $(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$ from the data represented through the GLC is less than what can be learnt from microdata then their respective estimated posterior distributions will differ even when the ABC approximation to the likelihood is exact (i.e., when $\tau \to 0$).

Together, the parametric framework for income distributions under 'missing rich' issues developed in the previous section along with the Bayesian empirical strategy

Algorithm 2: An AM ABC (***ABC-AM***) algorithm.

Initialization:

Set $\Sigma^{(0)}$, J_0 , M , τ

Until $K_\tau^{gauss}(\tilde{\varepsilon}^{(0)}) > 0$:

  Sample $(\tilde{\boldsymbol{\theta}}^{(0)}, \tilde{\boldsymbol{\eta}}^{(0)}, \tilde{\boldsymbol{\nu}}^{(0)}) \equiv \tilde{\boldsymbol{\phi}}^{(0)}$ from $p(\tilde{\boldsymbol{\phi}})$

  Generate $\{GLC_k^{Obs}(\tilde{\boldsymbol{\phi}}^{(0)})\}_{k=1}^N$ by simulating from $f_y^{Obs}(.;\tilde{\boldsymbol{\phi}}^{(0)})$

  Generate $\tilde{\varepsilon}^{(0)} = d(\{GLC_k^{Obs}\}_{k=1}^N, \{GLC_k^{Obs}(\tilde{\boldsymbol{\phi}}^{(0)})\}_{k=1}^N)$

Sampling: **for** $j = 1, ..., J$ **do**

Sample $\{\tilde{\boldsymbol{\phi}}^{(m)}\}_{m=1}^M \sim g_{\Sigma^{(j-1)}}(\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}^{(j-1)})$ from the candidate $g_{\Sigma^{(j-1)}}$

Generate $\{GLC_k^{Obs}(\tilde{\boldsymbol{\phi}}^{(m)})\}_{k=1}^N$ by simulating from $f_y^{Obs}(.;\tilde{\boldsymbol{\phi}}^{(m)})$, $m = 1, ..., M$

Generate $\tilde{\varepsilon}^{(j)} = \min_{m \in \{1,...,M\}} d(\{GLC_k^{Obs}\}_{k=1}^N, \{GLC_k^{Obs}(\tilde{\boldsymbol{\phi}}^{(m)})\}_{k=1}^N)$ and candidate $\tilde{\boldsymbol{\phi}}^{(j)} = \arg\min_{\tilde{\boldsymbol{\phi}}^{(m)}} \{d(\{GLC_k^{Obs}\}_{k=1}^N, \{GLC_k^{Obs}(\tilde{\boldsymbol{\phi}}^{(m)})\}_{k=1}^N)\}_{m=1}^M$

Accept and store $(\tilde{\boldsymbol{\phi}}^{(j)}, \tilde{\varepsilon}^{(j)})$ with probability:

$$\rho^{(j)} = \min\left\{1, \frac{K_\tau^{gauss}(\tilde{\varepsilon}^{(j)}) \times p(\tilde{\boldsymbol{\phi}}^{(j)}) \times g_{\Sigma^{(j-1)}}(\tilde{\boldsymbol{\phi}}^{(j-1)}, \tilde{\boldsymbol{\phi}}^{(j)})}{K_\tau^{gauss}(\tilde{\varepsilon}^{(j-1)}) \times p(\tilde{\boldsymbol{\phi}}^{(j-1)}) \times g_{\Sigma^{(j-1)}}(\tilde{\boldsymbol{\phi}}^{(j)}, \tilde{\boldsymbol{\phi}}^{(j-1)})}\right\}$$

▷ e.g., if $u^{(j)} \leq \rho^{(j)}$ where $u^{(j)}$ is a draw from a $Uniform(0,1)$ distribution

otherwise store $(\tilde{\boldsymbol{\phi}}^{(j)}, \tilde{\varepsilon}^{(j)}) = (\tilde{\boldsymbol{\phi}}^{(j-1)}, \tilde{\varepsilon}^{(j-1)})$

if $j > J_0$ **then**

Update $\Sigma^{(j)}$

end

end

presented in this section allow for a broad range of applications. The following section illustrates some of the main income distribution analysis possible under this approach.

5 Applications and examples

5.1 Applications on simulated data

Simulated-data applications can give insight on the performance of the ABC approach in making inference on $(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu})$ in a controlled setting exploiting a model following (1). Consider a hypothetical population's income distribution following a GB2 distribution $y_i \sim f_{\boldsymbol{y}}^{GB2}(.;\boldsymbol{\theta}) \equiv GB2(\alpha, \beta, p, q)$, with parameters α, p, and q ruling the shape of the distribution and β ruling the scale[11]. Typically, these parameters are the focus of the analysis of the income distribution. However, if the available data $\boldsymbol{y}^{Obs}$ is presumably affected by any of the 'missing rich' forms considered in the previous sections, additional parameters ruling assumed parametric forms for these issues must also be introduced into the analysis.

Assume that microdata samples from this population's income distribution may be jointly affected by high-income under-reporting following an LPU scheme with parameters $(\bar{p}, \delta)$ and high-income non-response following a right-truncation scheme with parameter t where $t \gg \bar{p}$. Noting that under this joint scheme[12]:

$$\frac{\partial m^{-1}(y_i^{Obs};\bar{p},\delta)}{\partial y_i^{Obs}} = \frac{1}{1-\delta \times \mathbf{1}(y_i^{Obs} > F_{\boldsymbol{y}}^{-1;GB2}(\bar{p};\alpha,\beta,p,q))}$$

and

$$\int f_{\boldsymbol{y}}^{GB2}(m^{-1}(y_i^{Obs};\bar{p},\delta);\alpha,\beta,p,q) \times \varphi(m^{-1}(y_i^{Obs};\bar{p},\delta);t) \times \left(\frac{\partial m^{-1}(y_i^{Obs};\bar{p},\delta)}{\partial y_i^{Obs}}\right) dy^{Obs} = t$$

, a model for the observable data $\boldsymbol{y}^{Obs}$ can be obtained applying (1)[13]:

$$\begin{aligned} f_{\boldsymbol{y}^{Obs}}(y_i^{Obs};\boldsymbol{\theta},\bar{p},\delta,t) &= \frac{f_{\boldsymbol{y}}^{GB2}\left(m^{-1}(y_i^{Obs};\bar{p},\delta);\boldsymbol{\theta}\right) \times \left(\frac{\partial m^{-1}(y_i^{Obs};\bar{p},\delta)}{\partial y_i^{Obs}}\right) \times \varphi(m^{-1}(y_i^{Obs};\bar{p},\delta);t)}{\int f_{\boldsymbol{y}}^{GB2}(m^{-1}(y_i^{Obs};\bar{p},\delta);\boldsymbol{\theta}) \times \left(\frac{\partial m^{-1}(y_i^{Obs};\bar{p},\delta)}{\partial y_i^{Obs}}\right) \times \varphi(m^{-1}(y_i^{Obs};\bar{p},\delta);t) dy^{Obs}} \\ &= \frac{f_{\boldsymbol{y}}^{GB2}\left(y_i^{Obs} + \mathbf{1}(y_i^{Obs} > F_{\boldsymbol{y}}^{-1;GB2}(\bar{p};\boldsymbol{\theta})) \times \left(\frac{\delta(y_i^{Obs} - F_{\boldsymbol{y}}^{-1;GB2}(\bar{p};\boldsymbol{\theta}))}{1-\delta}\right);\boldsymbol{\theta}\right) \times \mathbf{1}(y_i^{Obs} \leq (1-\delta)F_{\boldsymbol{y}}^{-1;GB2}(t;\boldsymbol{\theta}) + \delta F_{\boldsymbol{y}}^{-1;GB2}(\bar{p};\boldsymbol{\theta}))}{t \times (1-\delta \times \mathbf{1}(y_i^{Obs} > F_{\boldsymbol{y}}^{-1;GB2}(\bar{p};\boldsymbol{\theta})))} \end{aligned} \tag{3}$$

Equation (3) expands the GB2 distribution to allow for LPU (whenever $\bar{p} \ll t$ and $\delta > 0$) and for non-response in the form of a right-truncation (whenever $\bar{p} \ll t < 1$). For

[11] The GB2 distribution $GB2(\alpha, \beta, p, q)$ has pdf:

$$y_i \sim f_y^{GB2}(y_i|\alpha,\beta,p,q) = \frac{\alpha y_i^{\alpha p-1}}{\beta^{\alpha p} B(p,q)\left(1+\left(\frac{y_i}{\beta}\right)^{\alpha}\right)^{p+q}}\ ,\ (y_i,\beta,\alpha,p,q) \in \mathbb{R}_+^5$$

where $B(p,q)$ denotes the Beta function, defined as $B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt$. See Chotikapanich et al. (2018) for other parametric expressions under this distribution such as the Gini coefficient.

[12] In what follows, $F^{-1;GB2}(.;\alpha,\beta,p,q)$ denotes the quantile function of the GB2 distribution.

[13] For simplicity of notation, $\boldsymbol{\theta} = (\alpha,\beta,p,q)$ in what follows.

illustratory purposes, a first experiment of interest consists in estimating the posterior distribution $\pi(\boldsymbol{\theta}, \bar{p}, \delta, t|\boldsymbol{y}^{Obs})$ using the (***ABC-AM***) algorithm through this model over a sample of N simulated incomes $\boldsymbol{y}^{Obs} = \{y_i^{Obs}\}_{i=1}^{N}$. In particular, this exercise is most interesting when the simulated data is effectively affected by LPU and right-truncation forms of 'missing rich' jointly.

Benchmark parameter values can be set to $\left(\alpha, \frac{\beta}{1000}, p, q\right) = (2.3, 10, 1.75, 1.25)$ and $(\bar{p}, \delta, t) = (.5, .15, .99)$ in this interest[14]. These correspond to a population income distribution with an average income of 15054 and a Gini coefficient of 0.348. Data simulated under this setting corresponds to a sample from a GB2 distribution which starts being affected by LPU above the median with a slope of $\delta = .15$ and which contains no observations for units above the .99-th population's income distribution percentile.

Data can be simulated in this specific case by sampling $\frac{N}{1+t}$ incomes from the GB2 distribution and applying the LPU and right-truncation transformations under the benchmark values. This yields a single random sample of N observed incomes. The sample used in this exercise was generated in this way, for a hypothetical population of 10000 units (i.e., $N = 9900$). Figure 3 below illustrates how a sample generated under this setting relates to the theoretical observed incomes' distribution $f_{\boldsymbol{y}^{Obs}}$ under (3) and to the respective complete population's $f_{\boldsymbol{y}}^{GB2}$ income distribution.

Several conditions must be considered in elliciting a joint prior probability distribution for the model parameters $p(\boldsymbol{\theta}, \bar{p}, \delta, t)$. Firstly, this joint prior distribution can be set as the product of several marginal prior distributions:

$$p(\boldsymbol{\theta}, \bar{p}, \delta, t) = p(\alpha) \times p\left(\frac{\beta}{1000}\right) \times p(p) \times p(q) \times p(\bar{p}) \times p(\delta) \times p(t)$$

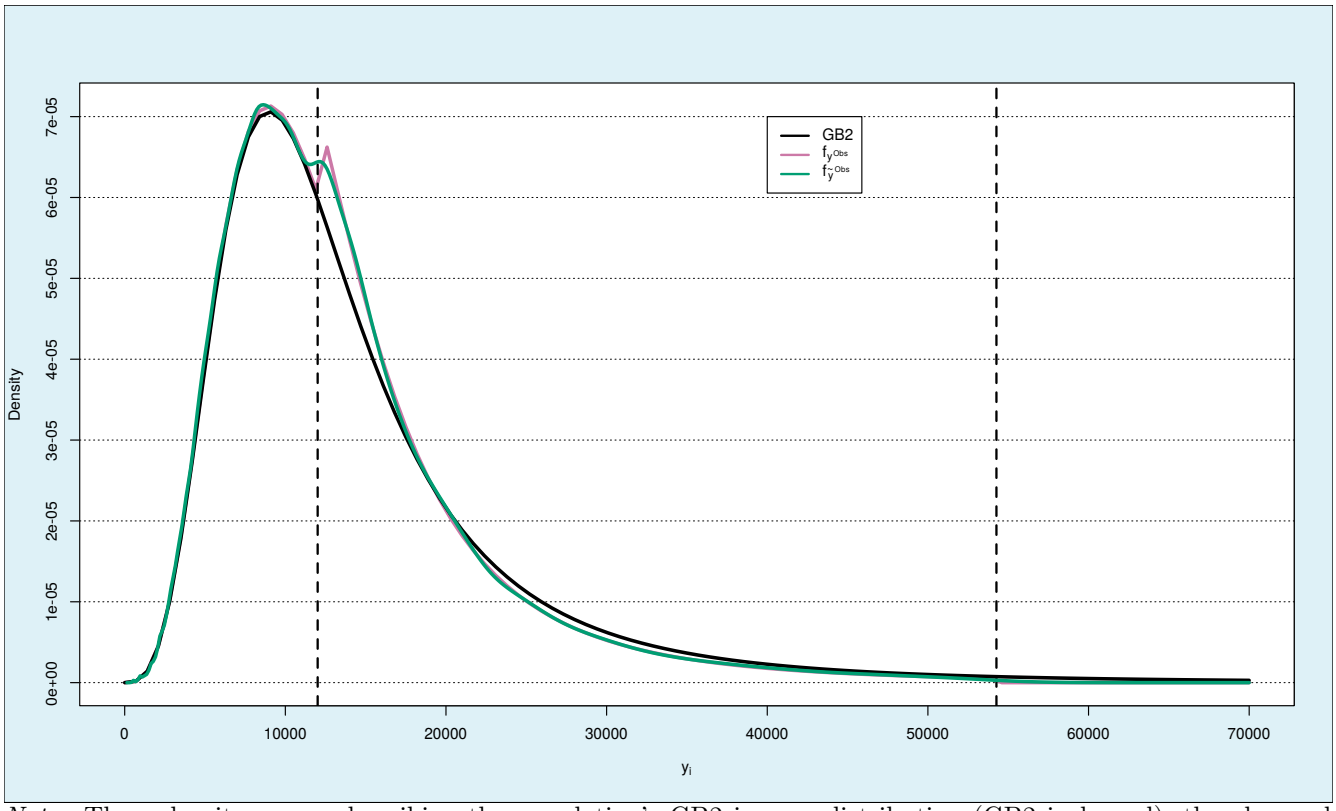
Secondly, given the high flexibility of the GB2 distribution it is possible to represent virtually any specific case of this distribution in a constrained range of parameter values. In this sense, the marginal prior distributions for the GB2 parameters were set as follows:

$$\begin{aligned} \alpha &\sim p(\alpha) \equiv Gamma(1, 1) \\ \frac{\beta}{1000} &\sim p\left(\frac{\beta}{1000}\right) \equiv Gamma(5, 2) \\ p &\sim p(p) \equiv Gamma(1, 1) \\ q &\sim p(q) \equiv Gamma(1, 1) \end{aligned}$$

. This amounts to prior beliefs on the shape parameters α, p, and q following a right-skewed Gamma distribution with mode at the value 1 and to prior beliefs on $\frac{\beta}{1000}$ following another right-skewed Gamma distribution with mode approximately at the value 8.

[14]The benchmark value for β being 10000, it is here scaled by 1000 in the interest of numerical stability when applying (***ABC-AM***).

Figure 3: Population, theoretical, and sample densities for model (3)



Note: Three density curves describing the population's GB2 income distribution (GB2 in legend), the observed incomes density function following (3) ($f_{y^{Obs}}$ in legend), and kernel density estimate from the estimating sample generated from this model ($N = 9900$) ($f_{\tilde{y}^{Obs}}$ in legend). Benchmark parameter values taken as $\left(\alpha, \frac{\beta}{1000}, p, q\right) = (2.3, 10, 1.75, 1.25)$ and $(\bar{p}, \delta, t) = (.5, .15, .99)$, with the left-most vertical dashed line representing the population's $\bar{p}$-th income percentile and the right-most vertical dashed line representing the right-truncation point.

Thirdly, reflecting a strong prior belief on the presence 'missing rich' issues in the data, the $(\bar{p}, \delta, t)$ parameters were given the following prior distributions:

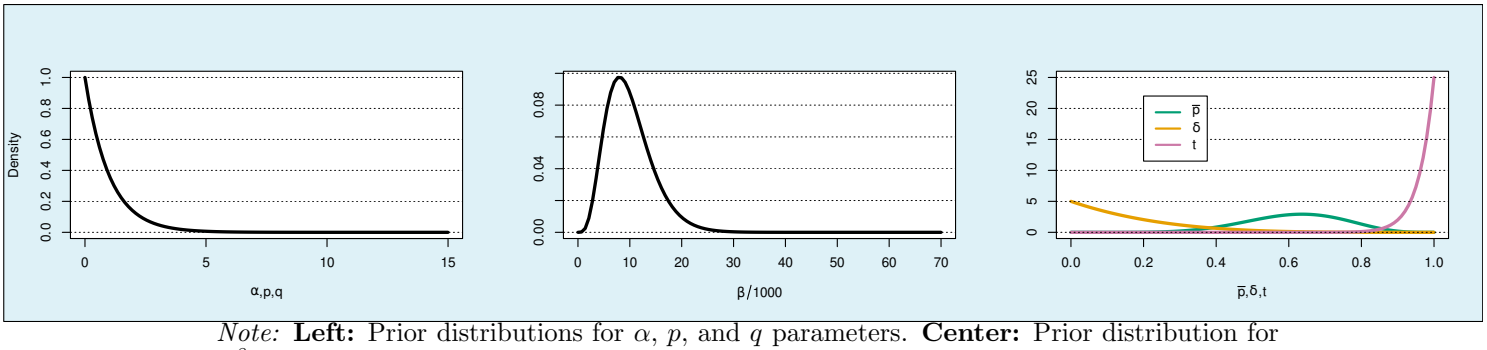
$$\begin{aligned} \bar{p} &\sim p(\bar{p}) \equiv Beta(8, 5) \\ \delta &\sim p(\delta) \equiv Beta(1, 5) \\ (1 - t) &\sim p(1 - t) \equiv Beta(1, 25) \end{aligned}$$

These reflect empirically-relevant values for the literature using right-truncation forms for non-response (e.g., see Jorda and Niño-Zarazúa 2019) and that exploring high-income under-reporting in survey data (e.g., see Flachaire et al. 2022). Importantly, these prior beliefs also give considerable probability to the 'complete data' scenario where no under-reporting or non-response issues affect the sample. This is, it is also made likely *a priori* that the observed income distribution may be correctly represented by a single GB2 distribution without introducing 'missing rich' phenomena.

Finally, several constraints may be imposed on the ellicited joint prior distribution to further constrain the parameter space. Imposing restrictions for finite variance on the GB2 income distribution amounts to giving 0 prior probability to parameter values with $\alpha < \frac{2}{q}$ and $\alpha < -\frac{1}{p}$. Additionally, because under-reported incomes have no relevance

if they correspond to a true income above the truncation point, the restriction $t > \bar{p}$ is imposed[15]. Figure 4 below summarizes these ellicited prior distributions for each of the model's parameters.

Figure 4: Prior distributions



Note: **Left:** Prior distributions for α, p, and q parameters. **Center:** Prior distribution for $\frac{\beta}{1000}$. **Right:** Prior distributions for $\bar{p}$, δ, and t parameters.

Three central scenarios are explored in applying the (***ABC-AM***) algorithm. Firstly, to evidence the biases that these forms of 'missing rich' induce if the issue is not taken into consideration, a simple GB2 distribution is fit to the data. A second scenario consists of estimating the income distribution parameters (α, β, p, q) under (3) conditional on fixing the correction quantities $(\boldsymbol{\eta}, \boldsymbol{\nu}) = (\bar{p}, \delta, t)$ at their true values. Finally, a third scenario consists of estimating all parameters in (3) elliciting prior uncertainty in $(\boldsymbol{\eta}, \boldsymbol{\nu})$. In all cases, the algorithm is set with parameters $\tau = 25$, $M = 10$, $J_0 = 15000$, and $\Sigma^{(0)} = diag(.01, .1, .01, .01, .01, .01, .01)$, and $J = 250000$ MCMC samples are obtained, discarding the initial 50000 considered as the burn-in period where the algorithm's adaptive terms are being calibrated. For computational ease, the simulated data taken as estimating sample was summarized by its GLC computed at sample centiles $\{GLC_k^{Obs}\}_{k=1}^{100}$.

Figure 5 below illustrates the goodness-of-fit of the resulting estimates for all three scenarios explored, along with the estimating sample $\{GLC_k^{Obs}\}_{k=1}^{100}$. The posterior predictive distributions computed following $\{GLC_k^{Obs}(\tilde{\boldsymbol{\phi}}^{(j)})\}_{k=1}^{100}$, $j = 1, ..., 200000$ graphically match the estimating data very closely for all scenarios, with very narrow 95% highest posterior density intervals[16] reflecting little uncertainty around the predicted values. This close fit to the data even holds at the top 10% of the income distribution, where the 'missing rich' issues are progressively more present. An additional implication of these results is that both the sample Gini coefficient and the sample mean income

[15]Formally, these restrictions impose the following joint prior distribution:

$$p(\boldsymbol{\theta}, \bar{p}, \delta, t) = p(\alpha) \times p\left(\frac{\beta}{1000}\right) \times p(p) \times p(q) \times p(\bar{p}) \times p(\delta) \times p(t) \times \prod_{i=1}^{3} C_{(i)}(\boldsymbol{\theta}, \bar{p}, \delta, t)$$

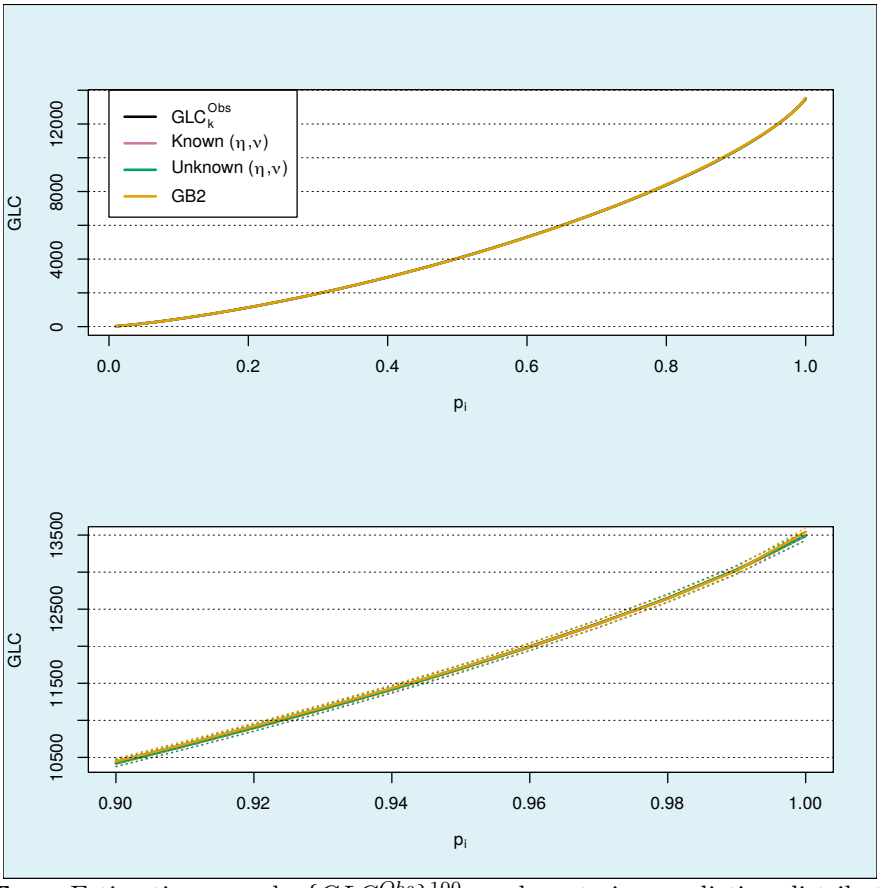
with

$$\begin{cases} C_{(1)}(\boldsymbol{\theta}, \bar{p}, \delta, t) = \mathbf{1}\left(\alpha > \frac{2}{q}\right) \\ C_{(2)}(\boldsymbol{\theta}, \bar{p}, \delta, t) = \mathbf{1}\left(\alpha > -\frac{1}{p}\right) \\ C_{(3)}(\boldsymbol{\theta}, \bar{p}, \delta, t) = \mathbf{1}\left(t > \bar{p}\right) \end{cases}$$

[16]Highest posterior density intervals are computed in what follows as the narrowest interval within the estimated posterior distribution accumulating 95% of the mass of the distribution.

are accurately fit in all three scenarios, as both are quantities of the GLC. Because the GLC is determined jointly by all parameters in the model, however, this overall good fit may hide important biases or differences in the estimated posterior distributions for each parameter individually.

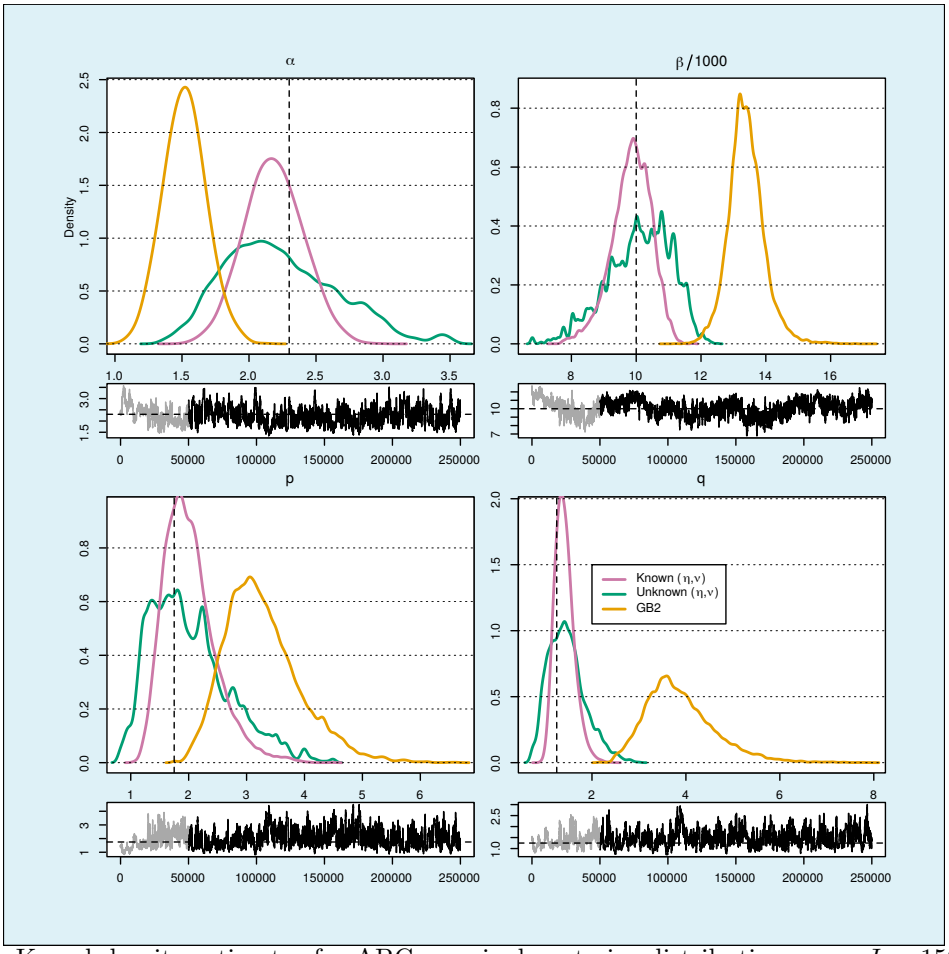
Figure 5: Posterior predictive estimates of the GLC



Note: **Top:** Estimating sample $\{GLC_k^{Obs}\}_{k=1}^{100}$ and posterior predictive distributions of $\{GLC_k^{Obs}(\tilde{\phi})\}_{k=1}^{100}$. Estimates separately obtained without 'missing rich' corrections (GB2, in legend), conditional on the true $(\bar{p}, \delta, t) = (.5, .15, .99)$ correction parameters, and with prior uncertainty on these. 95% highest posterior density interval bounds in respective dashed lines. **Bottom:** Same as top, focusing only on top 10% of the income distribution.

Estimated ABC marginal posterior distributions for each of the income distribution parameters in (3) are summarized in figure 6 below. As a first observation, all scenarios yield posterior distribution estimates which significantly differ from the ellicited prior distributions, effectively updating these prior beliefs. The most relevant result is the strong bias for all parameters affecting the estimates obtained without considering 'missing rich' issues. In contrast, both scenarios correcting for these issues yield estimated posterior distributions centered at their true value. Additionally, introducing uncertainty on the $(\bar{p}, \delta, t)$ parameters yields posterior distributions for the income distribution parameters which reflect higher uncertainty than the respective estimates obtained under the known true values for these.

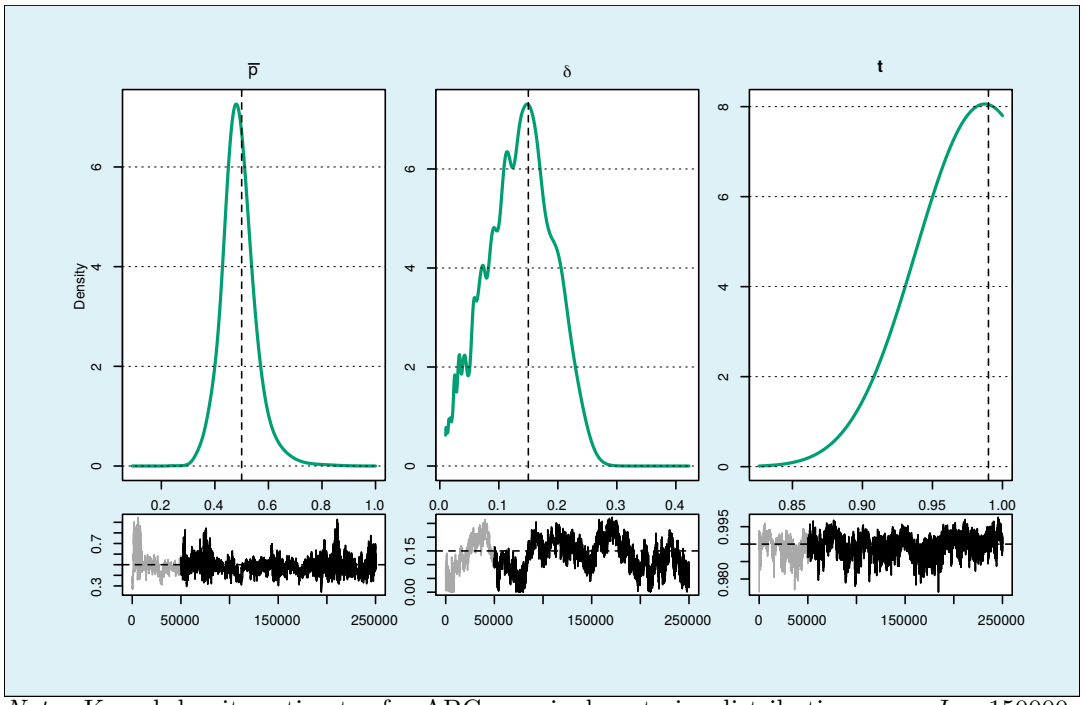
Figure 6: Estimated ABC marginal posterior distributions for income distribution parameters



Note: Kernel density estimates for ABC marginal posterior distributions over $J = 150000$ samples $\{\left(\alpha^{(j)}, \frac{\beta^{(j)}}{1000}, p^{(j)}, q^{(j)}\right)\}_{j=1}^{150000}$ from the (***ABC-AM***) algorithm after 50000 burn-in samples. Estimates separately obtained without 'missing rich' corrections (GB2, in legend), conditional on the true $(\bar{p}, \delta, t) = (.5, .15, .99)$ correction parameters, and with prior uncertainty on these. Traceplots of the underlying MCMC samples for estimates with uncertainty on $(\boldsymbol{\eta}, \boldsymbol{\nu})$ below, with burn-in period in gray. True parameter values in dashed black lines.

For the scenario where $(\bar{p}, \delta, t)$ are also to be inferred from the data, figure 7 below summarizes the estimated posterior distributions for these parameters. The estimates showcase a significant update of the ellicited prior distributions, with posterior distributions centered at the true values for these correcting quantities. This result provides support of the ABC approach as a fruitful empirical strategy for parametric inference on income distributions through data affected by 'missing rich' issues of uncertain magnitude.

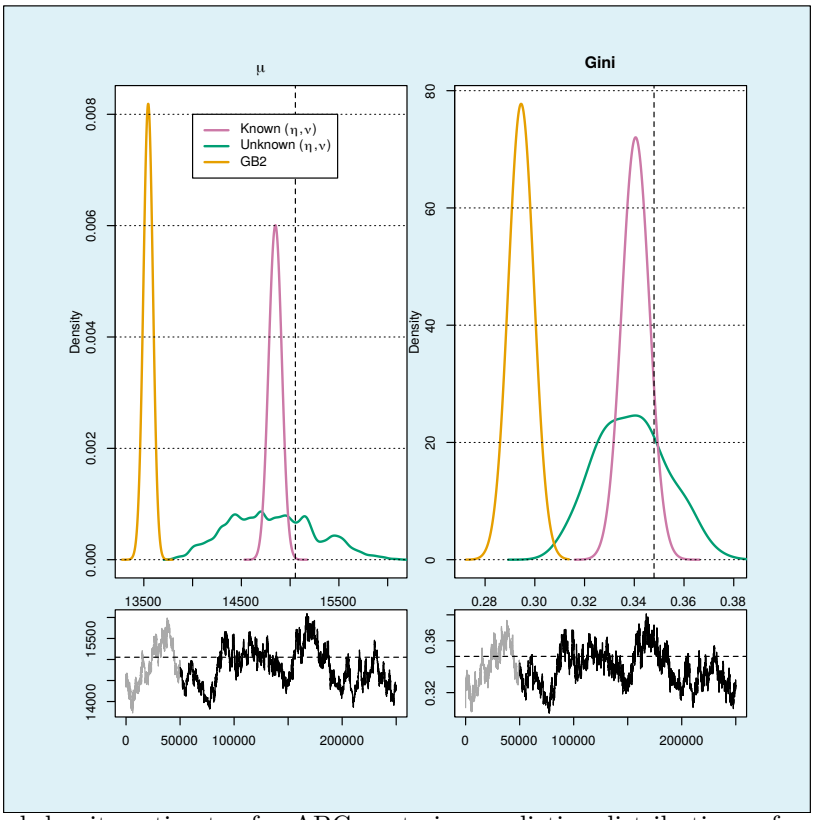
Figure 7: Estimated ABC marginal posterior distributions for 'missing rich' parameters



Note: Kernel density estimates for ABC marginal posterior distributions over $J = 150000$ samples $\{(\bar{p}^{(j)}, \delta^{(j)}, t^{(j)})\}_{j=1}^{150000}$ from the (***ABC-AM***) algorithm after 50000 burn-in samples. Traceplots of the underlying MCMC samples for estimates with uncertainty on $(\boldsymbol{\eta}, \boldsymbol{\nu})$ below, with burn-in period in gray. True parameter values in dashed black lines.

One additional illustration that this exercise provides concerns the impact of correcting for 'missing rich' issues in making inference on income growth and inequality at the population level. Figure 8 presents the posterior predictive distributions of the population's mean income and Gini coefficient, both determined by the (α, β, p, q) coefficients alone. These distributions evidence a significant under-estimation of both growth (through the mean income) and inequality (through the Gini coefficient) when the 'missing rich' issues are neglected. Only the scenarios estimated under corrections through $(\bar{p}, \delta, t)$ achieve estimates closely reproducing the true value at the population level. Despite all three estimates closely reproducing the estimating sample GLC, the bias affecting parameter estimates when 'missing rich' issues are not considered make inference on the population's income distribution invalid.

Figure 8: Posterior predictive estimates of population mean income and Gini coefficient



Note: Kernel density estimates for ABC posterior predictive distributions of population's mean income (μ) and Gini coefficient over $J = 150000$ samples $\{(\bar{p}^{(j)}, \delta^{(j)}, t^{(j)})\}_{j=1}^{150000}$ from the (***ABC-AM***) algorithm after 50000 burn-in samples. Estimates separately obtained without 'missing rich' corrections (GB2, in legend), conditional on the true $(\bar{p}, \delta, t) = (.5, .15, .99)$ correction parameters, and with prior uncertainty on these. Traceplots of the underlying MCMC samples for estimates with uncertainty on $(\boldsymbol{\eta}, \boldsymbol{\nu})$ below, with burn-in period in gray. True values in dashed black lines.

5.2 Real data applications:

As an illustratory example on real data, the European Union's Statistics on Income and Living Conditions (EU-SILC) provide an interesting household survey setting. EU-SILC data provides information on people and households within the EU representative at the country level, covering most countries in the EU yearly since 2005, and under a common framework defining the exact definitions of incomes and populations to be surveyed.

Although several calibrations are done over EU-SILC samples and sample weights for enforcing population representativeness on several dimensions, these are not done on income variables. Recent analysis have explored 'missing rich' phenomena on EU-SILC data (e.g., see Hlasny and Verme 2018,Bartels et al. 2019, Angel et al. 2019, Carranza et al. 2022, Ederer et al. 2022), suggesting this issue to be present with different magnitudes in all countries and periods.

A key income variable in income distribution analysis on EU-SILC data is household disposable income under the OECD-modified equivalence scale[17] (`HX090`). This considers all gross household incomes in the data net of regular taxes on wealth, regular inter-household transfers paid, and regular taxes on income and social insurance contributions.

Although this aggregate variable includes definitions of income variables that are common to all countries covered by EU-SILC, the sources from which the data is obtained differ across cross-sectional waves of data and countries. In particular, while some countries rely entirely on survey responses to measure these income variables, other countries source these variable entirely or partially from administrative registers. These differences in sources across waves and countries introduce large heterogeneities in the quality of the data in terms of under-reporting as registers are considered more reliable than survey responses. Additionally, the rising use of register sources determines that for some countries some of the waves of data are sourced from surveys and other waves are sourced from register. This can produce trends in the observed income distributions across waves without necessarily reflecting trends of the true population's income distribution.

Information on non-response rates for each country and wave of EU-SILC is publicly available through the corresponding quality reports published by the European Commission. Household non-response rates, in particular, can be informative about the overall degree of non-response affecting an observed distribution of household incomes. These rates are computed from a country-level household reponse rate, which is the product of address contact rates (i.e., the share of households in the sampling frame that were successfully contacted) and household response rates (i.e., the share of households in the sampling frame that completed their survey after being successfully contacted).

Table 1 below summarizes EU-SILC samples for five selected countries (Austria, Germany, France, Spain, and Italy) and for the 2005, 2007, 2011, and 2016 waves. The

[17]The OECD-modified equivalence scale computes a household's size `HX050` as:

`HX050` = 1+0.5×(#household members aged 14 and over −1)+0.3×(#household members aged 13 or less)

mean and Gini coefficient for household disposable income distributions summarize the observable trends in growth and inequality across countries and waves. Because these distributions are presumable affected by 'missing rich' issues, these values serve as a lower bound estimate for the corresponding population's mean and Gini coefficient. With many heterogeneities, all countries experienced mean income growth and, with the exception of Italy, income inequality increased from 2005 to 2016. Finally, overall household non-response rates show large disparities across countries and years in terms of levels and trends, illustrating possible heterogeneities in the incidence of this issue on the respective observed income distributions.

Table 1: EU-SILC sample descriptives for selected countries

Country	**Wave**	N	**Household non-response rate**	μ^{Obs}	**Gini**
Austria (AT)	2005	5146	0.38	20212.24	0.27
	2007	6805	0.22	20405.17	0.28
	2011	6182	0.23	23948.16	0.29
	2016	5992	0.27	26274.72	0.28
Germany (DE)	2005	13078	0.35	18078.73	0.27
	2007	14047	–	20084.84	0.31
	2011	13473	0.21	21047.33	0.30
	2016	13260	0.23	23424.24	0.31
France (FR)	2005	9745	0.16	18237.49	0.29
	2007	10485	0.14	18423.25	0.27
	2011	11348	0.18	23934.22	0.31
	2016	11446	0.17	25788.05	0.30
Spain (ES)	2005	12865	0.28	12289.05	0.33
	2007	12234	0.23	13520.56	0.32
	2011	12993	0.22	16535.78	0.33
	2016	14168	0.20	16151.14	0.34
Italy (IT)	2005	21874	0.15	16648.63	0.33
	2007	20809	0.14	17422.81	0.32
	2011	19234	0.25	18491.99	0.32
	2016	20966	0.21	18839.71	0.32

Source: Own calculations from EU-SILC.
Note: EU-SILC samples for Austria (AT), Germany (DE), France (FR), Spain (ES), and Italy (IT) from 2005, 2007, 2011, and 2016 waves. Only considers households with reported household disposable income (`HX090`) of at least 1 euro. Household non-reponse rates as reported in the publicly-available quality reports for each wave. Weighted-sample estimates of mean incomes (μ^{Obs}) and Gini coefficients.

Under the same settings as in the simulated data application, model (3) can be fit to the EU-SILC samples through the (***ABC-AM***) algorithm. Figure 9 below summarizes the resulting estimates in terms of goodness-of-fit to the weighted-sample mean income and Gini coefficient, as well as the estimated population values of these. As a first observation, the obtained results reproduce very accurately the levels and trends of mean incomes and Gini coefficients for all waves and countries considered.

Concerning the estimates of the population income distribution, a first observed result is these do not match their sample counterparts for any case. This suggests that

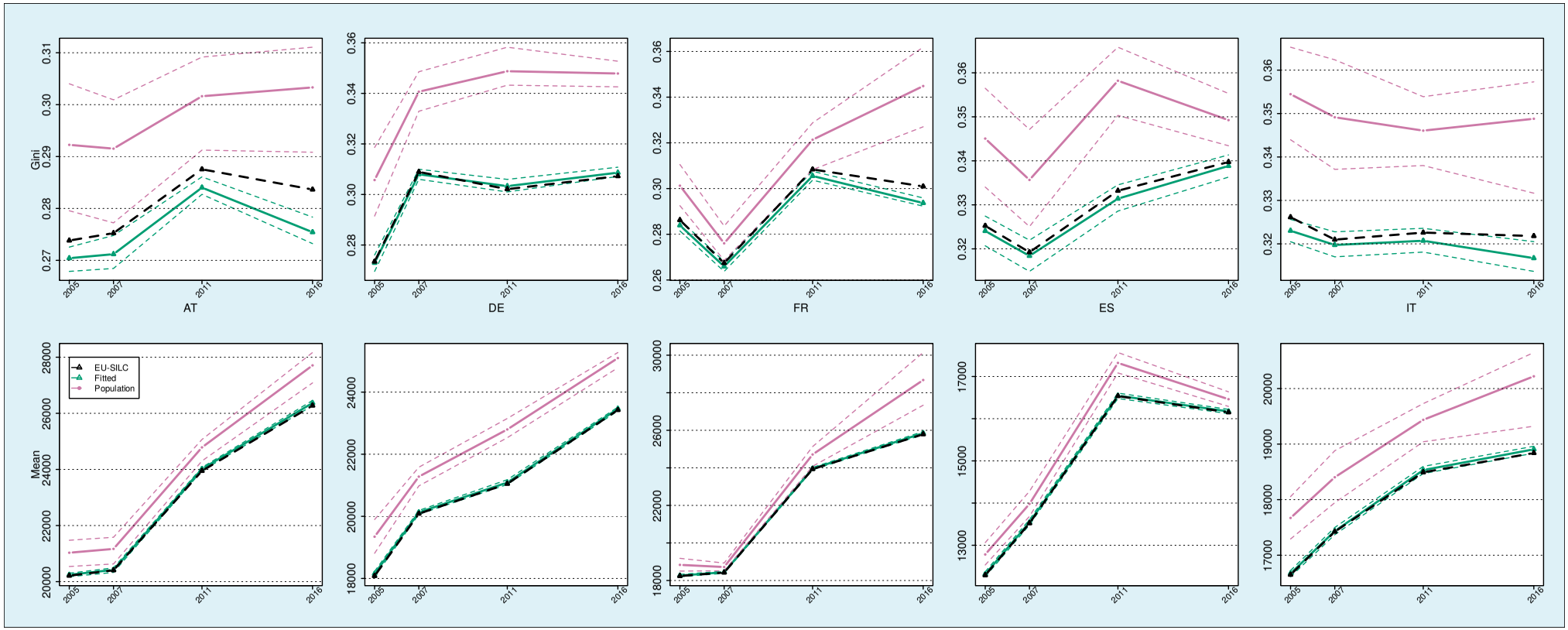
'missing rich' corrections are required in order to reproduce the observed distributions so accurately. Consequently, the estimated GB2 population income distribution parameters imply levels of income growth and inequality above those observed in the data. These population level estimates reproduce similar changes in mean incomes across waves as those in the observed samples, with some increase in the uncertainty around these quantities for the latter years in the case of France and Italy. Population estimates that reproduce the same dynamic of mean incomes as in the observed sample can be indicative of the total mass of incomes affecting the data due to 'missing rich' issues changing very little across waves.

As summarized by the Gini coefficient, the income inequality dynamics implied by these estimates at the population level pose some contrasts with their sample counterpart. In all cases inequality is estimated to be higher at the population level than what sample estimates suggest. The non-overlap of the computed credible intervals across population- and sample-level estimates of the Gini provide strong evidence of this. Another observation is that the uncertainty around the population Gini is relatively stable across EU-SILC waves for all countries, with strong heterogeneities across countries. This uncertainty is in some cases high enough to make inequality increases and decreases across time equally likely, as is the case for the Austrian Gini between the 2011 and 2016 waves. In other cases, however, the estimates provide clearer evidence of significant increases in income inequality across periods, as can be seen for France and Germany across 2005 and 2016 waves.

Conditional on the assumed forms for 'missing rich' issues in these applications, the estimated parameters can suggest margins where the representativeness of the data changes across waves. The detailed estimates in Appendix B suggest that the right-truncation parameter t introducing high-income non-response in the model (3) is estimated to be within the top .1% of the population income distribution in all waves and countries. This can suggest that significant non-response issues are mostly concentrated on households within this income group.

Concerning under-reporting issues, these estimates suggest strong heterogeneities in the share of the population affected by progressive under-reporting and the progressiveness of under-reporting across waves and countries. As quantified by the estimates for the $\bar{p}$ and δ parameters these are estimated to range from .5401 to .8525, and from .0779 to .3997 respectively across the selected EU-SILC samples. Taken together with the estimates for t, these results illustrate that reproducing the observed income distributions accurately under model (3) always requires jointly correcting for progressive under-reporting of incomes above the median and for right-truncation non-responses somewhere within the top .1% of the income distribution.

Figure 9: Posterior predictive estimates of mean income and Gini coefficients - EU-SILC countries



Source: Own calculations from EU-SILC.
Note: Posterior mean estimates of mean income and Gini coefficients for Austria (AT), Germany (DE), France (FR), Spain (ES), and Italy (IT) from 2005, 2007, 2011, and 2016 EU-SILC waves. Respective 95% highest posterior density intervals in dashed lines. In black, weighted-sample estimates of mean incomes and Gini coefficients. Only considers households with reported household disposable income (`HX090`) of at least 1 euro. Posterior distribution estimates obtained under model (3) over $J = 150000$ samples $\{\left(\boldsymbol{\theta}^{(j)}, \boldsymbol{\eta}^{(j)}, \boldsymbol{\nu}^{(j)}\right)\}_{j=1}^{150000}$ from the (***ABC-AM***) algorithm after 50000 burn-in samples.

6 Concluding remarks

Building on previous 'missing rich' correction methods explored in the literature, a new framework for parametric income distributions is proposed. This framework directly deals with these corrections based on expanding pre-existing parametric distributions with functional forms for both reporting and non-response mechanisms. As a model for data on incomes presumably affected by 'missing rich' this framework allows for devising empirical strategies to infer jointly features of the associated population's income distribution and features of the 'missing rich' issues in the data.

In dealing with the several constraints that must be faced by such an empirical strategy, the ABC approach is proposed as a fruitful method. This Bayesian estimator allows for updating prior uncertainty on the true population income distribution and the often uncertain 'missing rich' quantities affecting the observed data. This is achieved by attempting to reproduce the observed GLC with simulated incomes from the specified model.

The illustrative applications presented in this paper evidence some of the virtues of the ABC approach for inference on the parameters of GB2 income distributions under data affected by 'missing rich'. In a simulated-data setting, the analysis illustrates the several biases affecting inference on a population's income growth and inequality when 'missing rich' issues affecting the estimating data are neglected. This experiment also suggests the ABC approach to be fruitful in learning about uncertain 'missing rich' quantities given informative prior beliefs about these.

Applications to cross-sectional EU-SILC data on incomes give insight on the suitability of the framework in a typical household survey setting. The resulting estimates imply that reproducing the observed income distributions accurately requires considering both high-income under-reporting and non-response in all settings analysed. The analysis also illustrates how inference on population income distributions can be made under uncertain 'missing rich' quantities, uncovering contrasts between some of the observed trends at the sample level and those inferred for the respective population.

Further work seeking to understand the pitfalls of the proposed approach could explore further setups in terms of the specified 'missing rich' parametric forms. In principle, if these forms are capable of representing similar patterns of under-reporting and non-response, then estimates obtained under them should yield very similar results at the population level.

Another possible line of analysis for future work concerns exploring this empirical strategy in a linked-data setting. If comparing survey-sourced incomes and register-sourced incomes evidences some form of progressive under-reporting and high-income non-response, then estimates obtained under this framework using the survey data alone should be found to reproduce 'missing rich' patterns consistent with these.

In the specific case of the EU-SILC, an additional direction for future work concerns integrating available external information on the representativeness of the observed income distribution into the prior beliefs for the 'missing rich' quantities. In particular,

household non-response rates and information about the specific sources for the observed incomes for a given wave of data and country can be exploited in setting up informative prior probabilities. This could help in accounting for possible artificial trends arising from changes in the sampling or income sources and not from actual changes in the population's income distribution.

Finally, a possible extension of this framework involves making inference on income distributions of populations defined at aggregate levels such as regions or the globe. Taking the mixture of all countries' income distributions, estimates obtained accounting for 'missing rich' issues at the country level can be used to study patterns of income growth and distribution on aggregate levels.

References

Alvaredo, F. (2011). A note on the relationship between top income shares and the Gini coefficient. *Economics Letters*, 110(3):274–277.

Alvaredo, F., Chancel, L., Piketty, T., Saez, E., and Zucman, G. (2018). The elephant curve of global inequality and growth. In *AEA Papers and Proceedings*, volume 108, pages 103–08.

Angel, S., Disslbacher, F., Humer, S., and Schnetzer, M. (2019). What did you really earn last year?: explaining measurement error in survey income data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4):1411–1437.

Atkinson, A. B. and Piketty, T. (2007). *Top incomes over the twentieth century: a contrast between continental european and english-speaking countries*. OUP Oxford.

Atkinson, A. B., Piketty, T., and Saez, E. (2011). Top incomes in the long run of history. *Journal of economic literature*, 49(1):3–71.

Bartels, C., Metzing, M., et al. (2019). An integrated approach for a top-corrected income distribution. *Journal of Economic Inequality*, 17(2):125–143.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):235–269.

Berthet, P., Fort, J.-C., and Klein, T. (2020). A central limit theorem for Wasserstein type distances between two distinct univariate distributions. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pages 954–982. Institut Henri Poincaré.

Blanchet, T., Flores, I., and Morgan, M. (2018). The weight of the rich: Improving surveys using tax data. *WID. world Working Paper*, 2018/12.

Blanchet, T., Flores, I., and Morgan, M. (2022). The weight of the rich: improving surveys using tax data. *The Journal of Economic Inequality*, 20(1):119–150.

Bollinger, C. R., Hirsch, B. T., Hokayem, C. M., and Ziliak, J. P. (2019). Trouble in the tails? what we know about earnings nonresponse 30 years after Lillard, Smith, and Welch. *Journal of Political Economy*, 127(5):2143–2185.

Bourguignon, F. (2018). Simple adjustments of observed distributions for missing income and missing people. *The Journal of Economic Inequality*, 16(2):171–188.

Brunori, P., Salas-Rojo, P., and Verme, P. (2022). Estimating inequality with missing incomes. *ECINEQ Working Paper*, 616.

Burdín, G., Esponda, F., and Vigorito, A. (2014). Inequality and top incomes in Uruguay: a comparison between household surveys and income tax micro-data. *World Top Incomes Database Working Paper*, 1.

Burkhauser, R. V., Hérault, N., Jenkins, S. P., and Wilkins, R. (2017). Top incomes and inequality in the UK: reconciling estimates from household survey and tax return data. *Oxford Economic Papers*, 70(2):301–326.

Bustos, A. (2015). Estimation of the distribution of income from survey data, adjusting for compatibility with other sources. *Statistical Journal of the IAOS*, 31(4):565–577.

Carranza, R., Morgan, M., and Nolan, B. (2022). Top income adjustments and inequality: An investigation of the EU-SILC. *Review of Income and Wealth*.

Charpentier, A. and Flachaire, E. (2022). Pareto models for top incomes and wealth. *The Journal of Economic Inequality*, 20(1):1–25.

Chesher, A. and Schluter, C. (2002). Welfare measurement and measurement error. *The Review of Economic Studies*, 69(2):357–378.

Chotikapanich, D., Griffiths, W., Hajargasht, G., Karunarathne, W., and Rao, D. (2018). Using the GB2 income distribution. *Econometrics*, 6(2):21.

Chotikapanich, D. and Griffiths, W. E. (2000). Applications: posterior distributions for the Gini coefficient using grouped data. *Australian & New Zealand Journal of Statistics*, 42(4):383–392.

Chotikapanich, D. and Griffiths, W. E. (2008). Estimating income distributions using a mixture of Gamma densities. *Modeling Income Distributions and Lorenz Curves*, 5:285.

Clarté, G., Robert, C. P., Ryder, R. J., and Stoehr, J. (2021). Componentwise approximate Bayesian computation via Gibbs-like steps. *Biometrika*, 108(3):591–607.

Darvas, Z. (2019). Global interpersonal income inequality decline: The role of China and India. *World Development*, 121:16–32.

Deaton, A. (2005). Measuring poverty in a growing world (or measuring growth in a poor world). *Review of Economics and statistics*, 87(1):1–19.

Drovandi, C. and Frazier, D. T. (2021). A comparison of likelihood-free methods with and without summary statistics. *arXiv preprint arXiv:2103.02407*.

Eckernkemper, T. and Gribisch, B. (2021). Classical and bayesian inference for income distributions using grouped data. *Oxford Bulletin of Economics and Statistics*, 83(1):32–65.

Ederer, S., Ćetković, P., Humer, S., Jestl, S., and List, E. (2022). Distributional national accounts (DINA) with household survey data: Methodology and results for European countries. *Review of Income and Wealth*, 68(3):667–688.

Flachaire, E., Lustig, N., and Vigorito, A. (2022). Underreporting of top incomes and inequality: A comparison of correction methods using simulations and linked survey and tax data. *Review of Income and Wealth*.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press.

Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient Metropolis jumping rules. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics V. Proceedings of the Fifth Valencia International Meeting*. Oxford University Press.

Gottschalk, P. and Huynh, M. (2010). Are earnings inequality and mobility overstated? the impact of nonclassical measurement error. *The Review of Economics and Statistics*, 92(2):302–315.

Graf, M. and Nedyalkova, D. (2014). Modeling of income and indicators of poverty and social exclusion using the generalized beta distribution of the second kind. *Review of Income and Wealth*, 60(4):821–842.

Greenlees, J. S., Reece, W. S., and Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77(378):251–261.

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science*, 20(2):111–140.

Haario, H., Saksman, E., Tamminen, J., et al. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Hartley, M. J. and Revankar, N. S. (1974). On the estimation of the Pareto law from under-reported data. *Journal of Econometrics*, 2(4):327–341.

Higgins, S., Lustig, N., Vigorito, A., et al. (2018). The rich underreport their income: Assessing bias in inequality estimates and correction methods using linked survey and tax data. Tulane Economics Working Paper Series 1808, Tulane University.

Hinkley, D. V. and Revankar, N. S. (1977). Estimation of the Pareto law from underreported data: A further analysis. *Journal of Econometrics*, 5(1):1–11.

Hlasny, V. (2020). Nonresponse bias in inequality measurement: Cross-country analysis using Luxembourg Income Study surveys. *Social Science Quarterly*, 101(2):712–731.

Hlasny, V. and Verme, P. (2015). Top incomes and the measurement of inequality: A comparative analysis of correction methods using Egyptian, EU, and US survey data. In *ECINEQ Conference Paper*, volume 145.

Hlasny, V. and Verme, P. (2018). Top incomes and inequality measurement: a comparative analysis of correction methods using the EU SILC data. *Econometrics*, 6(2):30.

Hlasny, V. and Verme, P. (2022). The impact of top incomes biases on the measurement of inequality in the United States. *Oxford Bulletin of Economics and Statistics*, 84(4):749–788.

Hurst, E., Li, G., and Pugsley, B. (2014). Are household surveys like tax forms? evidence from income underreporting of the self-employed. *Review of economics and statistics*, 96(1):19–33.

Jenkins, S. P. (1995). Did the middle class shrink during the 1980s? UK evidence from kernel density estimates. *Economics letters*, 49(4):407–413.

Jenkins, S. P. (2009). Distributionally-sensitive inequality indices and the GB2 income distribution. *Review of Income and Wealth*, 55(2):392–398.

Jenkins, S. P. (2017). Pareto models, top incomes and recent trends in UK income inequality. *Economica*, 84(334):261–289.

Jorda, V. and Niño-Zarazúa, M. (2019). Global inequality: How large is the effect of top incomes? *World Development*, 123:104593.

Jorda, V., Sarabia, J. M., and Jäntti, M. (2021). Inequality measurement with grouped data: parametric and non-parametric methods. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(3):964–984.

Kantorovich, L. V. (1939). The mathematical method of production planning and organization. *Management Science*, 6(4):363–422.

Kobayashi, G. and Kakamu, K. (2019). Approximate Bayesian computation for Lorenz curves from grouped data. *Computational Statistics*, 34(1):253–279.

Korinek, A., Mistiaen, J. A., and Ravallion, M. (2007). An econometric method of correcting for unit nonresponse bias in surveys. *Journal of Econometrics*, 136(1):213–235.

Krishnaji, N. (1970). Characterization of the Pareto distribution through a model of underreported incomes. *Econometrica*, pages 251–255.

Lakner, C. and Milanovic, B. (2016). Global income distribution: From the fall of the Berlin wall to the great recession. *The World Bank Economic Review*, 30(2):203–232.

Leigh, A. (2009). Top incomes. In *The Oxford handbook of economic inequality*. New York: Oxford University Press.

Liu, J. S., Liang, F., and Wong, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134.

Lustig, N. (2019). Measuring the distribution of household income, consumption and wealth: State of play and measurement challenges. In Stiglitz, J. E., Fitoussi, J.-P., and Durand, M., editors, *For Good Measure: An Agenda for Moving Beyond GDP*. The New Press.

Lustig, N. (2020). The missing rich in household surveys: Causes and correction approaches. *ECINEQ Working Paper No. 2020-520*.

Lyssiotou, P., Pashardes, P., and Stengos, T. (2004). Estimates of the black economy based on consumer demand approaches. *The Economic Journal*, 114(497):622–640.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica*, 52(3):647–663.

Nandram, B. and Choi, J. W. (2002). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97(458):381–388.

Peters, G. and Sisson, S. A. (2006). Bayesian inference, monte carlo sampling and operational risk. *Journal of Operational Risk*, 1(3):27–50.

Pissarides, C. A. and Weber, G. (1989). An expenditure-based estimate of Britain's black economy. *Journal of public economics*, 39(1):17–32.

Ransom, M. R. and Cramer, J. S. (1983). Income distribution functions with disturbances. *European Economic Review*, 22(3):363–372.

Ratmann, O. R. (2010). *Approximate Bayesian Computation under model uncertainty*. PhD thesis, Imperial College London (University of London).

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72(359):538–543.

Shorrocks, A. F. (1983). Ranking income distributions. *Economica*, 50(197):3–17.

Silva, M. (2023). Parametric estimation of income distributions using grouped data: an Approximate Bayesian computation approach. *Aix-Marseille School of Economics (AMSE) Working Paper*, 2023(10).

Van Praag, B., Hagenaars, A., and van Eck, W. (1983). The influence of classification and observation errors on the measurement of income inequality. *Econometrica*, pages 1093–1108.

Appendix A

Comparing income distributions through the Wasserstein distance.

When microdata on a sample of incomes $\{y_{(i)}\}_{i=1}^{n}$ is available we can estimate empirical quantiles of the distribution and base the inference on a parametric model's parameters for the population income distribution around fitting the model at these quantiles. This is so because, quantiles are informative on both shape and scale of the distribution. This allows for an ABC approach without the need of summarizing the data through a small set of summary statistics.

The Wasserstein distance, originally developed in Kantorovich (1939), was recently explored for the purpose of summary-free ABC inference in Bernton et al. (2019) and Drovandi and Frazier (2021).

The Wasserstein distance between an income distribution f_y with quantile function F_y^{-1} and a parametric distribution model for this distribution $f_y(.;\theta)$ with quantile function $F_y^{-1}(.;\theta)$ follows:

$$W_p(f_y, f_y(.;\theta)) = \left(\int_0^1 d\{F_y^{-1}(\lambda), F_y^{-1}(\lambda;\theta)\}^p d\lambda \right)^{\frac{1}{p}}$$

. In the case of $p = 1$ and $d(x, y) = |x - y|$ this can be consistently estimated from the sample of incomes $\{y_{(i)}\}_{i=1}^{n}$ with empirical distribution $\hat{f}_y$ and a simulated sample of equal size from the model with empirical distribution $\hat{f}_y(.;\theta)$ as (e.g., Berthet et al. 2020):

$$W_1(\hat{f}_y, \hat{f}_y(.;\theta)) = \frac{1}{n} \sum_{i=1}^{n} |y_{(i)} - \tilde{y}_{(i)}|$$

This latter formulation can be interpreted as a metric comparing all sample order statistics (i.e., all sample quantiles). In essence, this metric estimates the average absolute difference between quantiles of the two distributions.

A metric $\rho(\hat{f}_y, \hat{f}_y(.;\theta))$ may be specified under a similar logic comparing the first-order differences of all empirical GLC coordinates (i.e., estimates of the quantiles by definition of the GLC) between these microdata samples instead of order statistics directly then this amounts simply to taking the Wasserstein distance as defined above:

$$\begin{aligned}
\rho(\hat{f}_y, \hat{f}_y(.;\theta)) &= \sum_{i=1}^{n} |(GLC(y_{(i)}) - GLC(y_{(i-1)})) - (GLC(\tilde{y}_{(i)}) - GLC(\tilde{y}_{(i-1)}))| \\
&= \sum_{i=1}^{n} \left| \left(\frac{y_{(i)}}{\sum_{i=1}^{n} y_{(i)}} \right) \times \left(\frac{\sum_{i=1}^{n} y_{(i)}}{n} \right) - \left(\frac{\tilde{y}_{(i)}}{\sum_{i=1}^{n} \tilde{y}_{(i)}} \right) \times \left(\frac{\sum_{i=1}^{n} \tilde{y}_{(i)}}{n} \right) \right| \\
&= \sum_{i=1}^{n} \frac{\left| y_{(i)} - \tilde{y}_{(i)} \right|}{n} \\
&= W_1(\hat{f}_y, \hat{f}_y(.;\theta))
\end{aligned}$$

This result supports the use of the Wasserstein-1 distance as a common unidimensional discrepancy allowing for ABC inference either with microdata or grouped data summarized through the GLC.

To compute the Wasserstein-1 distance on grouped data in the form of K groups' GLC coordinates (i.e., observed incomes are split into K segments with bounds $z_{(k)}$, $k = 1, ..., K$) we could have:

$$
\begin{aligned}
\rho(\hat{f}_y, \hat{f}_y(.;\theta)) &= \sum_{k=1}^{K} |(GLC(y_{(k)}) - GLC(y_{(k-1)})) - (GLC(\tilde{y}_{(k)}) - GLC(\tilde{y}_{(k-1)}))| \\
&= \sum_{k=1}^{K} \left| \left(\frac{\bar{y}_{(k)} \times n_{(k)}}{\sum_{k=1}^{K} \bar{y}_{(k)} \times n_{(k)}} \right) \times \left(\frac{\sum_{k=1}^{K} \bar{y}_{(k)} \times n_{(k)}}{K} \right) - \left(\frac{\bar{\tilde{y}}_{(k)} \times n_{(k)}}{\sum_{k=1}^{K} \bar{\tilde{y}}_{(k)} \times n_{(k)}} \right) \times \left(\frac{\sum_{k=1}^{K} \bar{\tilde{y}}_{(k)} \times n_{(k)}}{K} \right) \right| \\
&= \frac{1}{K} \sum_{k=1}^{K} \left| \bar{y}_{(k)} \times n_{(k)} - \bar{\tilde{y}}_{(k)} \times n_{(k)} \right| \\
&= \frac{1}{K} \sum_{k=1}^{K} \left| (\bar{y}_{(k)} - \bar{\tilde{y}}_{(k)}) \times n_{(k)} \right| \\
&= \frac{1}{K} \sum_{k=1}^{K} \left| (\bar{y}_{(k)} - \bar{\tilde{y}}_{(k)}) \times n_{(k)} \right| \\
&= \frac{1}{K} \sum_{k=1}^{K} \left| \sum_{i=1}^{n} (y_{(i)} - \tilde{y}_{(i)}) \times I(z_{(k)} \geq y_{(i)} \geq z_{(k-1)}) \times I(z_{(k)} \geq \tilde{y}_{(i)} \geq z_{(k-1)}) \right|
\end{aligned}
$$

which, in the trivial case of having $K = n$ groups (i.e., one observation per group) corresponds to the expression for this distance on microdata.

These results suggest that in the case of grouped data we can exploit the discrepancies between GLC curves through their first-order difference (i.e., through the approximation to the Wasserstein-1 distance).

Geometrically, the Wasserstein-1 distance computes the average absolute difference between the quantile functions of two distributions. When only grouped data is available, this average distance could be approximated by first computing the area between both empirical quantile curves within each interval of the grouped data, sum these areas across all intervals and divide by the number of intervals. The approximation comes the fact that in computing these areas the curves might cross within an interval and so we would have no way of accounting for those differences which counteract within the interval (i.e., the absolute value is applied at the interval level in the grouped data expression above). For a same population size n, however, the quality of the approximation always increases with K.

Having access to microdata allows a computationally-cheap alternative in which user-specified groups or bins can be defined for exploiting the grouped-data approximation to the Wasserstein-1 distance. For instance, instead of grouping the data on sample deciles, one could define broader groups for lower incomes and finer groups for higher incomes

allowing for a particularly stricter fit on the upper tail of the distribution.

Appendix B

EU-SILC application.

Table 2: ABC posterior distribution estimates for selected EU-SILC samples under (3)

Country	Wave	$\boldsymbol{\theta}$				$\boldsymbol{\eta}$		$\boldsymbol{\nu}$
		α	$\frac{\beta}{1000}$	p	q	$\bar{p}$	δ	t (%)
Austria (AT)	2005	4.4926 [3.4804;5.4754]	17.4491 [16.4642;18.3747]	0.7689 [0.5323;1.0381]	0.6995 [0.4948;0.887]	0.6481 [0.5588;0.7402]	0.1394 [0.0604;0.219]	99.9212 [99.8606;99.9774]
	2007	4.2329 [3.3264;5.6431]	17.497 [16.3194;18.3236]	0.8706 [0.5311;1.1924]	0.7848 [0.5149;0.9549]	0.6917 [0.5881;0.809]	0.1654 [0.0536;0.2412]	99.9987 [99.996;100]
	2011	4.4908 [3.6792;5.4883]	22.5277 [21.8214;23.2466]	0.6266 [0.4651;0.7952]	0.7225 [0.5291;0.8909]	0.6646 [0.5484;0.7882]	0.1145 [0.0464;0.1831]	99.9475 [99.9076;99.9806]
	2016	4.1521 [3.2317;5.3752]	23.2533 [21.3724;24.4954]	0.7947 [0.5194;1.1763]	0.76 [0.5472;0.9347]	0.6726 [0.6233;0.7313]	0.2252 [0.1611;0.2871]	99.9989 [99.9969;100]
Germany (DE)	2005	4.1816 [3.243;5.0834]	15.6159 [14.7938;16.395]	0.8099 [0.5691;1.076]	0.7151 [0.5477;0.8938]	0.5401 [0.4771;0.6122]	0.2318 [0.1642;0.304]	99.9992 [99.9974;100]
	2007	5.1701 [4.464;5.8984]	17.1657 [16.7115;17.6088]	0.5078 [0.4107;0.6081]	0.4937 [0.4123;0.5753]	0.7639 [0.7268;0.7959]	0.2838 [0.2277;0.3407]	99.999 [99.9966;100]
	2011	4.2773 [3.5422;5.1335]	18.1032 [17.45;18.8769]	0.6265 [0.4699;0.8126]	0.5993 [0.4394;0.7026]	0.7678 [0.7408;0.7945]	0.3997 [0.3482;0.4484]	99.9995 [99.9982;100]
	2016	4.3893 [3.8794;4.8718]	20.3698 [19.8252;20.9009]	0.5862 [0.4868;0.6795]	0.5884 [0.5204;0.6677]	0.773 [0.7527;0.7889]	0.3453 [0.3142;0.3776]	99.9995 [99.9985;100]
France (FR)	2005	2.9131 [2.3348;3.4636]	13.4956 [11.8554;14.9544]	1.6066 [1.0129;2.3224]	1.1055 [0.8676;1.3669]	0.6539 [0.573;0.726]	0.1383 [0.0741;0.2019]	99.9923 [99.9772;100]
	2007	3.9296 [3.2279;4.6694]	15.6607 [15.1527;16.1589]	0.9887 [0.7278;1.2634]	0.8818 [0.6438;1.1277]	0.8525 [0.7985;0.9022]	0.1446 [0.0458;0.249]	99.9892 [99.9691;100]
	2011	4.4173 [3.2817;5.1189]	17.8792 [16.9069;18.7775]	0.8398 [0.5991;1.1511]	0.6043 [0.464;0.748]	0.5919 [0.4136;0.716]	0.0779 [0.011;0.1361]	99.9853 [99.9702;99.9978]
	2016	5.4397 [2.9187;8.3704]	18.4294 [16.2994;20.3333]	0.9048 [0.3647;1.6694]	0.5048 [0.2796;0.7936]	0.5747 [0.4268;0.7176]	0.2963 [0.2481;0.3755]	99.9996 [99.9988;100]
Spain (ES)	2005	1.6968 [1.2934;2.1302]	11.0949 [9.2004;12.9569]	2.2101 [1.2547;3.3318]	2.3738 [1.4735;3.414]	0.7627 [0.7141;0.8138]	0.2002 [0.1207;0.2806]	99.9714 [99.9194;100]
	2007	1.6921 [1.2519;2.1322]	12.7219 [10.3523;14.8156]	2.2679 [1.2379;3.5877]	2.5442 [1.6421;3.6412]	0.7057 [0.637;0.7789]	0.1481 [0.0696;0.2237]	99.9733 [99.9272;100]
	2011	2.5284 [2.1636;2.9317]	14.8681 [13.9019;15.8051]	1.0773 [0.8016;1.3675]	1.2021 [0.9435;1.4661]	0.8243 [0.7939;0.856]	0.2938 [0.2314;0.3629]	99.9946 [99.983;100]
	2016	2.5723 [2.2547;2.9011]	17.8008 [16.8279;18.7658]	0.8678 [0.7037;1.0366]	1.4029 [1.1189;1.6883]	0.82 [0.7653;0.8743]	0.1248 [0.0639;0.191]	99.9909 [99.9731;100]
Italy (IT)	2005	3.4764 [2.8887;4.0841]	14.038 [13.2232;14.817]	0.7882 [0.5732;1.0109]	0.7659 [0.613;0.9184]	0.6818 [0.6362;0.7277]	0.2253 [0.165;0.2872]	99.9889 [99.9745;100]
	2007	2.9708 [2.3382;3.6742]	14.818 [13.5246;16.1084]	0.9931 [0.6399;1.4089]	0.972 [0.7172;1.2408]	0.6871 [0.6274;0.7446]	0.2037 [0.132;0.2773]	99.9596 [99.9203;99.9933]
	2011	4.1234 [3.4491;4.8556]	17.1416 [16.4914;17.757]	0.5719 [0.4288;0.7125]	0.6735 [0.5503;0.7979]	0.6788 [0.6229;0.7496]	0.1961 [0.1454;0.2515]	99.9985 [99.9946;100]
	2016	3.3105 [2.319;4.7316]	19.1832 [18.358;20.2492]	0.7476 [0.381;1.1051]	0.9856 [0.5918;1.3342]	0.5769 [0.5323;0.6323]	0.1915 [0.0879;0.2723]	99.9738 [99.9444;99.9994]

Source: Own calculations from EU-SILC.
Note: Mean of ABC marginal posterior distribution estimates for all parameters of model (3) over $J = 150000$ samples $\{\left(\boldsymbol{\theta}^{(j)}, \boldsymbol{\eta}^{(j)}, \boldsymbol{\nu}^{(j)}\right)\}_{j=1}^{150000}$ from the (***ABC-AM***) algorithm after 50000 burn-in samples. 95% highest posterior density intervals in brackets. Estimates from EU-SILC samples for Austria (AT), Germany (DE), France (FR), Spain (ES), and Italy (IT) from 2005, 2007, 2011, and 2016 waves. Only considers households with reported household disposable income (`HX090`) of at least 1 euro.